Элементы математической лингвистики

А. В. ГЛАДКИЙ
И. А. МЕЛЬЧУК

А. В. ГЛАДКИЙ, И. А. МЕЛЬЧУК

Элементы математической лингвистики



ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
Москва 1969

518
Г 52
УДК 519·95

Элементы математической лингвистики. Гладкий А. В., Мельчук И. А.

В книге дается характеристика задач и методов математической лингвистики, вводятся ее основные понятия, излагается важнейший и наиболее полно разработанный ее раздел — теория формальных порождающих грамматик Н. Хомского, а также дается краткий обзор других разделов и направлений математической лингвистики. Все вводимые в книге математические понятия разъясняются с привлечением большого числа лингвистических примеров. Специально рассматривается вопрос о путях приложения формальных грамматик к изучению естественных языков.

Книга рассчитана, во-первых, на широкий круг лингвистов, для которых она может служить введением в проблематику математической лингвистики, во-вторых, на математиков и вообще специалистов в области точных наук, интересующихся проблемами языка. Для них книга может быть популярным очерком основ новой математической дисциплины, хорошо освещающим ее содержательную сторону.

Книга может также служить пособием по курсу математической лингвистики в университетах и педагогических институтах. Страниц 192. Таблиц 5. Иллюстраций 21.

Алексей Всеволодович Гладкий, Игорь Александрович Мельчук

Элементы математической лингвистики

М., 1969 г., 192 стр. с илл.

Редакторы *Г. В. Вакуловская, О. С. Кулагина*

Техн. редактор *К. Ф. Брудно* Корректор *В. П. Горячева*

Сдано в набор 31/I 1969 г. Подписано к печати 29/V 1969 г. Бумага $84\times108^1/_{32}$
Физ. печ. л. 6. Условн. печ. л. 10,08. Уч.-изд. л. 9,63.
Тираж 13000 экз. Т-06949. Цена книги 61 коп. Заказ 1842.

Издательство «Наука»
Главная редакция физико-математической литературы
Москва, В-71, Ленинский проспект, 15

2-я типография издательства «Наука», Москва, Шубинский пер, 10.

2-2-3
69-69

Содержание

От авторов

В процессе работы над книгой мы неоднократно пользовались дружеской помощью и советами целого ряда лиц. Наиболее существенные замечания и предложения были высказаны Ф. А. Дрейзиным, Л. Н. Иорданской, Б. А. Трахтенбротом, А. И. Фетом. Кроме того, рукопись была прочитана Ю. Д. Апресяном, А. Я. Диковским, А. А. Зализняком, О. С. Кулагиной, З. К. Литвинцевой, Е. В. Падучевой, В. А. Успенским, А. А. Холодовичем, Ю. А. Шрейдером и другими, от которых мы также получили немало ценных соображений. Всем эти лицам мы выражаем нашу искреннюю признательность. Их советы мы постарались учесть, что, впрочем, удалось нам не в полной мере. Ответственность за те или иные недостатки книги лежит, разумеется, целиком на авторах.

Академгородок — Гольяново,
март 1964 г. — март 1967 г.

Введение

Цель настоящей книги — кратко охарактеризовать относительно недавно возникшую область исследований, известную под названием «математическая лингвистика». Сделать это, однако, непросто. Когда нужно ответить, например, на вопрос «Что такое дифференцирование?» или «Что такое топология?», речь идет лишь о том, чтобы найти максимально доступную и эффективную форму изложения вещей, прекрасно известных специалистам, понимаемых всеми ими одинаково, так сказать, канонизированных. Математику, взявшемуся объяснить неподготовленному читателю, что такое дифференцирование, разумеется, не придется сначала решать этот вопрос для себя: ответ он хорошо знает заранее и должен заботиться лишь о методической стороне изложения (что, конечно, и само по себе представляет задачу весьма нелегкую). В нашем случае дело обстоит иначе. Математическая лингвистика — это дисциплина совсем молодая, находящаяся в процессе становления и еще не успевшая обзавестись традициями. Общепринятой точки зрения на ее задачи, методы и тем более границы и рамки не существует. Поэтому мы были вынуждены в известной степени заново вырабатывать некоторую точку зрения на эту область в целом, выяснять ее общее строение, пытаться наметить ее статус. Тем не менее, данная книга не является оригинальным исследованием; она представляет собой популяризацию, но популяризацию не вполне обычную: в ее основу положены не более или менее традиционные концепции, в данном случае отсутствующие, а некоторая общая картина математической лингвистики, предлагаемая самими авторами. (Мы ни в коей мере не претендуем ни на окончательность,

ни на единственность этой картины. Возможны, конечно, и другие подходы, ср., например, Plath 1961; авторы, однако, воздерживаются от их обсуждения.)

Уже из самого названия математической лингвистики видна ее связь как с математикой, так и с лингвистикой (иначе — языковедением, или языкознанием, т. е. наукой о естественных языках). Поэтому *) авторы стремились сделать эту книгу «читабельной» и для математиков, и для лингвистов. При этом основной читатель, на которого ориентируются авторы,— лингвист. Такой выбор не случаен. Во-первых, математическая лингвистика, как она понимается в этой книге, есть м а т е м а т и ч е с к а я дисциплина; стало быть, если мы хотим, чтобы с ней могли знакомиться и лингвисты, и математики, надо рассчитывать изложение на лингвистов, как находящихся в худшем положении: все, что поймет здесь лингвист, поймет в большинстве случаев и математик, но не наоборот. Во-вторых, именно лингвист в первую очередь заинтересован в содержательной интерпретации и приложениях построений математической лингвистики.



В связи со сказанным авторы строили изложение так, чтобы от читателя формально не требовалось никаких специальных математических знаний — все используемые понятия, вплоть до самых простых, обязательно разъясняются. Тем не менее следует честно признать, что для подлинного понимания содержания настоящей книги крайне желательно свободное владение некоторыми простейшими, но вместе с тем фундаментальными математическими понятиями, например понятиями множества и функции, а также элементарными приемами математических рассуждений; в частности, полезно иметь привычку к таким понятиям, как необходимое условие, достаточное условие, доказательство по индукции и т. п. Повторяем: все это не является, строго говоря, необходимым; однако по существу без этого обойтись трудно, хотя и можно — ценой затраты гораздо бо́льших усилий

*) О знаках на полях — см. стр. 14.

(поскольку по ходу чтения книги читатель все равно должен будет овладеть всеми указанными понятиями).

Вообще следует подчеркнуть, что лингвист, желающий так или иначе применять математику в своей исследовательской деятельности, нуждается в первую очередь именно в общематематической культуре, в привычке к математическому мышлению, а не в техническом аппарате математики.

Для чтения книги полезно также знакомство с современной структурной лингвистикой (работы Л. Ельмслева, Л. Блумфилда, З. Харриса, Н. Трубецкого, Р. Якобсона и др. *)); прямого отношения к математической лингвистике соответствующие исследования не имеют, однако именно они во многом подготовили почву для ее возникновения, выдвинув ряд понятий и представлений (ср., например, идею непосредственно составляющих), образовавших содержательную базу для формальных моделей математической лингвистики.

Особое место здесь занимают работы Н. Хомского и его школы. В отличие от упомянутых выше эти работы либо непосредственно относятся к математической лингвистике, либо тесно с ней связаны по существу. Более того, именно они дали первоначальный толчок к ее развитию, что позволяет считать Н. Хомского основоположником этой новой дисциплины. Естественно поэтому, что значительная часть книги (§§ 2, 3, 4 и частично § 5) посвящена в основном изложению учения Н. Хомского о формальных порождающих грамматиках. Следовательно, знакомство с соответствующими работами будет чрезвычайно полезно. В первую очередь это относится к книге Н. Хомского «Синтаксические структуры» (Хомский 1962); см. также Хомский 1965а, Хомский 1965б, и Хомский — Миллер 1965; для математика можно рекомендовать еще Хомский 1966.

Однако сказанное выше о математических знаниях относится и к наличию специальной лингви-

*) Прекрасным введением в проблематику и методологию структурной лингвистики может служить книга Ю. Д. Апресяна (Апресян 1966). См. также Глисон 1959.

стической подготовки: строго говоря, оно также не является обязательным—все существенные для изложения понятия и факты, выходящие за рамки школьной грамматики, поясняются.

Итак, чего же авторы требуют от читателя? Формально — ничего, т. е. никаких математических или лингвистических знаний; по существу же предполагается некоторая математическая и лингвистическая грамотность. (Эта ситуация в книгах по математике довольно обычна: нередко сложные монографии, чтение которых требует больших усилий даже от специалиста, начинаются формально справедливым утверждением, что у читателя не предполагается никаких предварительных знаний.)

Теперь уточним, что авторы предлагают читателю. В книжке он найдет очерк основных идей и понятий математической лингвистики, который может служить лишь для предварительного и самого общего ознакомления с нею. Если же у него возникнет желание углубить свое знакомство с математической лингвистикой, например, с целью самому работать в этой области, то он должен будет обратиться к специальной литературе, часть которой указана в конце книги, на стр. 189—193.

Нам остается еще разъяснить здесь ту методологическую установку, на которую опирается все последующее изложение. Мы имеем в виду вопрос о правомерности существенного использования математики в такой сугубо гуманитарной области, как изучение человеческих языков. Если бы книга была рассчитана исключительно на математиков, подобное разъяснение было бы, по-видимому, излишним; однако для многих читателей-лингвистов (а может быть, и не только лингвистов) оно, как нам кажется, необходимо. Дело в том, что в лингвистике этот вопрос неоднократно и еще совсем недавно бурно дискутировался. При этом целый ряд лингвистов находит, что «математическая лингвистика» есть contradictio in adjecto, что язык и математика несовместимы и ничего общего между собой не имеют, что попытки внедрить математику в лингвистику ведут

к дегуманизации последней и тем самым к ее гибели как самостоятельной науки. Сторонники таких взглядов считают математическую лингвистику «скрещением псевдолингвистики с псевдоматематикой» (Абаев 1965, стр. 32).

Авторы данной книги стоят на прямо противоположных позициях. Есть, разумеется, немало плохих работ, о которых остроумно замечено, что они представляют собой попытку применить то немногое, что автор знает из математики, к тому немногому, что он знает из лингвистики (Ю. К. Щеглов, устно). Однако за эти работы математическая лингвистика ответственности не несет — они не имеют к ней никакого отношения. Что же касается существа дела, то мы полностью убеждены не только в возможности, но и в необходимости математического описания языковых явлений. Мы не можем давать здесь конкретные разъяснения по существу этого описания: подобные разъяснения и составляют содержание предлагаемой книги. Однако мы настаиваем на том, что любое научное описание (в том числе описание языка) должно быть логически последовательным (исключается пропуск существенных *) звеньев рассуждения), однозначным (исключаются формулировки, допускающие более одного понимания) и вполне эксплицитным (исключается «контрабандное» привлечение информации, не входящей в описание в явном виде). А такое описание и есть формальное, т. е.— при достаточно высоком уровне формализации — по существу математическое описание. Слово «формальный» не означает ничего, кроме как «логически последовательный + однозначный + абсолютно явный», так что формальное описание отнюдь не исключает обращения к содержанию, к смыслу (вполне возможно, в частности, формальное описание смысла, см. ниже, стр. 153 и далее). Стало быть, если формализацию отождествляют с дегуманизацией, видя в этой последней максимально возможное



*) То есть таких, на автоматическое восстановление которых читателем пишущий рассчитывать не может.

исключение из описания «человеческого фактора» (т. е. субъективного начала, неоговоренного обращения к интуиции и сообразительности читателя), то мы — за такую дегуманизацию. Подобное исключение «человеческого фактора» из описаний есть неотъемлемая часть любого научного метода: без этой «дегуманизации» никакая наука невозможна.



Сделаем два существенных разъяснения:

1. Разумеется, в процессе построения описания, в процессе научного творчества интуиция, сообразительность и т. п., т. е. «человеческий фактор», играют ведущую роль, и этого никто не станет отрицать. Речь идет об изгнании «человеческого фактора» из результатов исследования, из самих описаний.

2. Сказанное выше отнюдь не означает, будто мы признаем только формализованные («математизированные») лингвистические работы, отказывая в ценности всем прочим. Ничего подобного! Несомненно, что недостаточно четкое логически исследование, содержащее новые важные факты или идеи, может быть, вообще говоря, гораздо ценнее, чем безупречная формализация тривиальной истины. Однако столь же несомненно, что из двух работ, описывающих один и тот же круг фактов с одинаковой полнотой, бо́льшую ценность представляет та, в которой достигнута более высокая степень формализации.

Таким образом, мы исходим из тезиса о желательности дегуманизации лингвистики (в указанном выше смысле слова), и тем самым вопрос о возможности математической лингвистики решается утвердительно.

В связи с признанием необходимости формализации лингвистических описаний целесообразно отметить следующее.

Формальное описание любого объекта неизбежно связано со схематизацией и огрублением наблюдаемой картины *). Человеку, воспитанному в классиче-

*) Очевидно, что по мере совершенствования наших знаний об объекте огрубление становится все меньше и меньше, а описание все ближе и ближе к действительности. Процесс такого приближения является, по-видимому, бесконечным.

ских «гуманитарных» традициях, подобный подход может показаться порочным. Однако это единственно возможный путь научного познания. В самом деле, познать некоторый сложный объект не означает ничего другого, как установить закономерности его строения, т. е. выделить составляющие его простые компоненты и сформулировать правила, по которым эти компоненты соединяются между собой. В результате получается именно с х е м а изучаемого объекта.

При этом различные аспекты рассмотрения этого объекта обычно приходится разделять. В действительности все такие аспекты могут быть тесно связаны и сложным образом взаимодействовать друг с другом. Тем не менее как раз для того, чтобы точно описать их взаимосвязь, необходимо сначала изучить их по отдельности. Вообще, успех научного исследования зависит прежде всего от умения разбить сложную задачу на более простые (т. е. «схематизировать» ее) и выделить нужные аспекты анализа простой задачи. Разумеется, необходимо, чтобы при подобных упрощениях и огрублениях сохранялись все существенные — с точки зрения стоящей перед исследователем цели — свойства и особенности объекта. Лишь в той мере, в какой соблюдается это условие, формализация будет полезной и плодотворной.

Что же касается конкретных формальных описаний, предлагаемых математической лингвистикой и приводимых в этой книге, то вопрос об их содержательной ценности — вполне законный вопрос, но в каждом конкретном случае он должен решаться особо. Читатель, возможно, найдет, что в тех или иных формальных построениях упущены какие-либо существенные стороны моделируемых языковых явлений. Подобных критических замечаний относительно математической лингвистики, действительно, можно сделать немало. Однако, на наш взгляд, это означает только, что определенные модели нуждаются в усовершенствовании или исправлении, и нисколько не затрагивает самого принципа формализации,

который является основным лозунгом математической лингвистики (как и любой другой точной науки).

Понятие «формального» не является абсолютным: возможны различные степени, или уровни, формализации, между которыми нет резких границ. Разумно, однако, выделять два типа формализации: доматематический, когда используемые понятия в большей мере сохраняют индивидуальные, содержательные особенности конкретного объекта, и математический, предполагающий использование только абстрактных сущностей, заданных точными определениями (в этом случае для установления связей между абстрактными единицами и конкретными объектами нужна специальная интерпретация).

Математическая лингвистика имеет дело с формализацией второго типа. Однако поскольку формализация является всегда не самоцелью, а средством изучения конкретных явлений, то для всякой точной науки, а стало быть, и для математической лингвистики важны не только ее формальные понятия и утверждения, но и их интерпретация, т. е. то, как эти понятия и утверждения прилагаются к настоящим объектам.

В связи с этим авторы стремились насытить изложение содержательными примерами, показывающими, как введенные формальные понятия работают применительно к реальным языкам. При этом сложность естественного языка такова, что полное описание даже (казалось бы) достаточно простого его фрагмента оказывается слишком громоздким для настоящей книги. Кроме того, построение полных (формальных) описаний для тех или иных фрагментов естественного языка не относится к собственным задачам математической лингвистики: она должна лишь вырабатывать средства и методы для таких описаний, а их применение к языку входит в компетенцию лингвистики как таковой (о соотношении математической лингвистики и «просто» лингвистики см. Заключение, стр. 174—176). Поэтому наши примеры, как правило, являются фрагментарными и носят по преимуществу иллюстративный характер.

Наконец, мы хотели бы подчеркнуть, что предлагаемая книга не является ни систематическим изложением математической лингвистики, ни обзором основных работ и результатов в этой области. О некоторых важных направлениях в книге говорится лишь вскользь, а многие интересные проблемы (и относящиеся к ним результаты) вообще не упомянуты. *Авторы видели свою задачу лишь в том, чтобы дать читателю общее представление о математической лингвистике.*

В соответствии с этим авторы стремились свести к минимуму количество ссылок на литературу, выбирая в первую очередь наиболее доступные в нашей стране и наименее трудные по изложению работы, по возможности — на русском языке (для всех привлекаемых иностранных работ, изданных в русском переводе, указывается только перевод).

Таким образом, имеющийся в конце книги список упоминаемых работ никоим образом не претендует на полноту и не должен рассматриваться как самостоятельный библиографический указатель по математической лингвистике; кроме того, в список включен ряд работ, не имеющих прямого отношения к математической лингвистике, но оказавшихся необходимыми по ходу изложения.

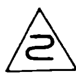
Поскольку книга, как отмечалось выше, предназначается не только (и даже не в первую очередь) для математиков, было сочтено целесообразным всемерно «облегчить» формулировки математических утверждений. Поэтому определения и теоремы во многих случаях излагаются в свободном стиле, вперемежку с содержательными пояснениями и т. п., т. е. в непривычной для математика форме (хотя и вполне строго по существу). Доказательства, как правило, не приводятся. Однако, в интересах читателя-математика, книга снабжена Приложением I (стр. 177—183), где собраны все математические утверждения, сформулированные вполне каноническим образом; там же даны ссылки на источники, где можно найти соответствующие доказательства.

В книге нет упражнений, однако, чтобы возместить их отсутствие, авторы в ряде случаев сознательно опускали некоторые звенья рассуждений (не очень существенные для главной линии изложения), не проводили полностью выкладок или не заканчивали разбора отдельных примеров в расчете на то, что читатель попытается заполнить эти лакуны сам. Все подобные случаи оговорены в тексте и отмечены специальным знаком на полях (см. ниже); иногда приводятся наводящие соображения. Выполнение этих «квазиупражнений» не является строго обязательным для понимания, хотя, разумеется, чем больше их будет сделано, тем лучше читатель овладеет материалом. Следует также иметь в виду, что задачи, оставленные на долю читателя, очень различны по трудности: от предельно простых до весьма нетривиальных.

С целью направлять внимание читателя авторы используют систему специальных знаков на полях:

NB — «обратить особое внимание»;

 — «очень важное звено рассуждений», «фундаментальное утверждение»;

 — «трудное место», «тонкое, но существенное различие», «парадоксальный (на первый взгляд) вывод»;

$\mathbf{T}_i$ — «утверждение, по существу представляющее собой формулировку математической теоремы» (индекс i означает номер соответствующей теоремы в Приложении I — сводке основных математических результатов по затронутой в тексте проблематике);

— «квазиупражнение» (см. выше).

§ 1. Содержание понятия «математическая лингвистика» (предварительные соображения)

Как обычно употребляют термин «математическая лингвистика»

Термин «математическая лингвистика» вошел в употребление в середине 50-х годов и к настоящему времени получил широкое распространение. Однако до сих пор разными людьми он понимается по-разному.

С одной стороны, слова «математическая лингвистика» чаще всего употребляются в очень широком и весьма расплывчатом смысле, а именно, их применяют к самым различным лингвистическим исследованиям, если в них хотя бы в незначительной степени используется математика или даже если малоискушенным читателям только кажется, будто она используется. Так, сюда относят и работы по созданию математических моделей языка с использованием аппарата алгебры, математической логики, теории алгоритмов; и работы, связанные со статистикой; и работы, где для формулирования тех или иных лингвистических положений привлекаются простые понятия и способы выражения (или обозначения), заимствованные из математики. Сюда же зачисляют все лингвистические работы, предполагающие применение вычислительных машин, в том числе работы по автоматическому переводу, даже если в них математика не используется ни по существу, ни по форме. Кроме того, под название «математическая лингвистика» часто подводятся всевозможные работы

прикладного характера и вообще лингвистические сочинения, носящие четко выраженный нетрадиционный характер.

Как следует употреблять этот термин

Подобное словоупотребление, т. е. применение термина «математическая лингвистика» для обозначения части лингвистики, представляется авторам неудачным. Оно создает ошибочное впечатление, будто существуют две разные лингвистики — одна принципиально не математическая, другая — особая, «математическая». В действительности же лингвистика есть единая наука со своими задачами и своим объектом, которая пользуется математическими методами там, где они нужны, и не пользуется ими там, где они не нужны.

В настоящей книге под «математической лингвистикой» понимается нечто совсем другое, а именно — определенный круг математических, по существу, работ, возникших из попыток строго описать факты естественных языков и содержащих результаты, которые могут оказаться полезными для лингвистики. При этом к указанному кругу не относятся те направления исследований, для которых язык является лишь одним из возможных приложений (в частности, работы чисто количественного характера, т. е. лингвистическая статистика и т. д.): для математической лингвистики в данном понимании характерно использование лишь тех математических методов, которые в определенном смысле (см. ниже) специфичны для языка как такового. Итак, математическая лингвистика есть математическая дисциплина, «обращенная» в сторону естественных языков и лингвистики.

Задача настоящей книги состоит в том, чтобы попытаться охарактеризовать объект и методы математической лингвистики, а также ее соотношение с лингвистикой, по возможности пользуясь при этом привычными для лингвистов представлениями,

Язык как отображение (функция)

Поскольку математическая лингвистика — это, как было уже сказано, математическая дисциплина, будет удобным начать изложение с рассмотрения следующей ситуации: математик, совершенно незнакомый с лингвистикой, наблюдает речевое поведение людей, т. е. функционирование языка, и пробует описать его; естественно, что полученное описание будет отражать присущий ему математический образ мышления и будет строиться с привлечением хотя бы простейших понятий математики. Это описание могло бы быть, например, таким. С одной стороны, наш математик видит, что содержанием речевой деятельности является передача различных желаний, чувств, представлений, мыслей и т. п. Все это он для краткости называет «планом содержания» (не пытаясь дать определение этому термину). С другой стороны, он видит, что средством передачи, или выражения, содержания служат последовательности физических сигналов (звуковых или графических), которые он называет «планом выражения». Для математика естественно представлять себе и план содержания, и план выражения как совокупности, или — пользуясь математической терминологией — множества некоторых элементов, которые он называет, допустим, соответственно «смыслами» и «текстами». Смыслы и тексты вовсе не обязаны быть простыми единицами: так, наблюдаемым текстом может быть слово, предложение, очень длинное высказывание и т. д., вплоть до целой книги (аналогично обстоит дело со смыслами). Далее, наш математик замечает, что между смыслами и текстами имеется соответствие: каждому смыслу отвечает более или менее определенная совокупность (множество) текстов, а каждому тексту — более или менее определенное множество смыслов. Правила, определяющие, какие тексты соответствуют каким смыслам, и образуют по существу то, что в обиходе принято называть языком. Математик же усмотрит в этой системе правил (т. е. в Языке!) частный случай важнейшего понятия своей

науки — о т о б р а ж е н и я, или ф у н к ц и и. Эта функция сопоставляет каждому смыслу некоторое (конечное) множество текстов, а именно, множество синонимичных текстов, несущих этот смысл (вообще говоря, один текст может соответствовать разным смыслам — омонимия). При этом наш математик заметит, что данная функция является, по-видимому, эффективно вычислимой (короче — эффективной) *); действительно, язык представляет собой некоторый регулярный способ эффективного получения текстов по заданным смыслам и обратно. Этот способ пока не известен математику-наблюдателю, но изучение речевого поведения людей приводит его к гипотезе, что такой способ, как-то «записанный» в мозгу носителей языка, безусловно имеется. При этом он знает, что изучением свойств эффективных функций занимается специальная математическая дисциплина — теория алгоритмов, являющаяся в свою очередь ответвлением математической логики. Если теперь наш математик захочет исследовать язык, то для него это будет означать исследование соответствующей функции. Он будет стремиться построить ее в явном виде как некоторую систему правил и одновременно начнет описывать ее свойства. Естественно, что он попытается обратиться к теории алгоритмов, как к источнику сведений относительно функций такого типа и методов их изучения.

Однако естественные языки образуют очень специфический, имеющий свои особые характеристики класс эффективных функций. И наш математик сразу же обнаружит, что сведений об эффективных функциях, почерпнутых из теории алгоритмов, для исследования естественных языков недостаточно. Следовательно, ему придется специально заняться изучением именно этих особых эффективных функций, тем самым развивая теорию алгоритмов в некотором новом,

*) Функцию $F(x)$ принято называть эффективно вычислимой, если для нее указан вполне определенный способ, позволяющий для любого значения x найти за конечное число шагов значение $F(x)$, т. е. если, грубо говоря, она может быть вычислена на машине.

нужном ему направлении. При этом он может столкнуться с необходимостью исследовать также и такие свойства интересующей его функции, которые не связаны с ее эффективностью (и поэтому, вообще говоря, не изучаются теорией алгоритмов) и носят очень общий абстрактный характер. Подобные свойства, т. е. свойства отображений, совершенно не зависящие от их конкретной природы (например, ассоциативность, коммутативность и т. п.), являются предметом алгебры. Стало быть, математик-лингвист должен будет воспользоваться еще и алгебраическими фактами и методами.

Приблизительно так и действовали, по всей вероятности, реальные математики, приступившие к систематическому исследованию естественного языка в 50-х годах (что стимулировалось, в частности, появлением ряда прикладных задач: автоматическая обработка языковой информации и т. д.). В результате их усилий сложилась специфическая математическая дисциплина со своей особой тематикой, исследования по которой ведутся в целом ряде стран и по которой публикуется большое количество работ. Эту дисциплину и предлагается называть математической лингвистикой.

Таким образом, математическая лингвистика — это область, которую можно рассматривать, с одной стороны, как специальную ветвь теории алгоритмов, а с другой стороны — как частный раздел алгебры, в связи с чем эту область иногда называют алгебраической лингвистикой. В некоторых работах по математической лингвистике преобладающую роль играет теория алгоритмов, в других — алгебра, а в третьих тесно переплетаются методы обеих дисциплин. Кроме того, нередко приходится пользоваться комбинаторными методами. Объектом же математической лингвистики являются функции (отображения) особого рода и различные возникающие в связи с ними абстрактные образования, в ряде существенных отношений сходные с естественными языками.

Необходимо подчеркнуть, что, как следует из сказанного, математическая лингвистика представ-

ляет собой в основном неколичественную дисциплину. Здесь, по-видимому, необходимо специальное разъяснение. В ряде наук основным методом описания свойств изучаемых объектов является установление соотношений между характеризующими эти объекты величинами. Например, в физике основные результаты — это количественные формулы; экспериментальная проверка физических утверждений сводится, как правило, к ряду измерений. Разумеется, количественные утверждения физически отражают глубокие качественные особенности физического мира; однако именно количественные соотношения являются основной формой описания этих особенностей.



Совсем иначе обстоит дело в лингвистике. Существенные характеристики языка не имеют количественной природы, т. е. не являются величинами; лингвистический эксперимент обычно не связан с измерениями. Старый тезис «Язык внеположен числу» имеет, при всей его расплывчатости, глубокое рациональное содержание. Вовсе не случайно, что основные достижения лингвистики, полученные за все время ее существования, не формулируются в виде количественных утверждений; вообще, принципиальная «неколичественность» типична для подавляющего большинства лингвистических работ. Такое положение связано, на наш взгляд, с природой самого объекта лингвистики — со свойствами естественного языка—и представляется поэтому вполне закономерным. Однако детальное обоснование высказанных соображений должно быть темой отдельного исследования *).

Положению о неколичественном характере математической лингвистики вовсе не противоречит тот

*) Вообще говоря, количественные утверждения о языке возможны (чаще всего эти утверждения относятся к распределению тех или иных величин, вроде длины фраз или слов, частоты слов того или иного типа и т. п.; ср., например, известный закон Ципфа). Мы, однако, полагаем, что в описании языка такие утверждения всегда носят периферийный характер.

факт, что в ней (равно как и в «обычной» лингвистике) в ряде случаев оказываются необходимыми те или иные вычисления: ср. использование выкладок для доказательства существования алгоритма (стр. 52—53), оценки сложности выводов в грамматиках (стр. 116—119), алгоритм классификации букв (стр. 171—173). Подобные обращения к количеству, сколь бы существенными они ни были, всегда играют подчиненную, вспомогательную роль — в том смысле, что их конечной целью является получение чисто качественных результатов, не представимых в виде количественных зависимостей. Так, хотя только что упомянутый алгоритм классификации букв основан на вычислениях, в результате его работы получается разделение букв на гласные и согласные, что само по себе с количеством никак не связано. Оценки сложности алгоритмов и выводов, которые имеют в математической лингвистике немалое значение (и, вероятно, в дальнейшем будут иметь значение еще более важное), нужны только для суждений об адекватности/неадекватности тех или иных моделей, т. е. опять-таки для сугубо неколичественных утверждений.

Существенная особенность вычислений, применяемых при изучении и описании естественных языков, состоит, по мнению авторов, в том, что такие вычисления во многих случаях должны относиться скорее не к речи, а к самому языку, т. е. к системе. Другими словами, наиболее плодотворными представляются не подсчеты типа «Сколько раз встречается в данных текстах данное слово, данная конструкция, слова данного класса и т. п.»: определение числа каких-либо объектов в большинстве интересных случаев выполняется не экспериментально (т. е. не прямым пересчетом этих объектов, например, в реальном тексте), а чисто дедуктивно (рассуждениями, исходящими из их абстрактных свойств).

Математическая лингвистика является неколичественной дисциплиной в той же степени, в какой признаются неколичественными современная алгебра или теория алгоритмов. В этих областях также

нередко привлекаются количественные соображения, которые тем не менее остаются в них на втором плане.

Мы остановились на вопросе о неколичественном характере математической лингвистики столь подробно только потому, что среди филологов существует убеждение о тождестве математической лингвистики со статистикой речи *). Несмотря на его распространенность, подобный взгляд фактически глубоко ошибочен и методически вреден.

*) В недавно вышедшем стабильном учебнике для вузов (Реформатский А. А., Введение в языковедение, М., 1967) читаем: «...Математическая лингвистика... является не особой лингвистикой, а лишь применением к языковым явлениям математических методов. Главным образом, это относится к речи, а не к языку, например применение теории вероятностей и математической статистики» (стр. 50).

§ 2. Формальные грамматики

О понятии формальной грамматики

Чтобы дать читателю более конкретное представление о математической лингвистике, мы остановимся на одном из ее разделов, который в настоящее время оказался наиболее разработанным, а именно, на так называемой *теории грамматик*. Под *грамматиками* в математической лингвистике понимаются некоторые специальные системы правил, задающие (или характеризующие) множества *цепочек* (конечных последовательностей) символов. Эти цепочки могут интерпретироваться как языковые объекты различных уровней, например как словоформы (цепочки морф *)), словосочетания и предложения (цепочки словоформ) и т. п.

Таким образом, грамматики математической лингвистики — *формальные грамматики* — имеют дело с абстракциями, возникающими путем обобщения таких обычных лингвистических понятий, как словоформа, словосочетание, предложение.

Поясним, что имеется в виду, когда говорится «формальные грамматики задают множества цепочек». Из данного набора символов (обозначающих, например, все словоформы русского языка) можно строить какие угодно цепочки; некоторые из этих цепочек естественно считать правильными, или допустимыми (например, грамматически правильные предложения — *Сосны шумят на ветру* или *Радости свистят на меху*), а другие — неправильными, или недопу-

---

*) Смысл понятия «морфа» можно приблизительно разъяснить для нелингвиста следующим образом: морфы — это минимальные осмысленные (или имеющие самостоятельную функцию) части словоформ, например *пере* + *доказ* + *ыва* + + *ем* + *ому* (плюс обозначает границы между морфами).

стимыми (**Ветру на шумят меху*; звездочкой отмечаются неправильные фразы, словоформы и т. п.). Формальная грамматика задает (характеризует) правильные цепочки, если имеет место одно из двух:

1) либо для любой предъявленной цепочки грамматика умеет решить, является эта цепочка правильной или нет, и в случае положительного ответа дать указания о строении этой цепочки;

2) либо грамматика умеет построить любую правильную цепочку, давая при этом указания о ее строении, и не строит ни одной неправильной цепочки.

В первом случае формальная грамматика называется **распознающей**, во втором — **порождающей** *).

Выше указывалось, что с математической точки зрения язык представляет собой некоторую эффективную функцию. Эта функция имеет чрезвычайно сложное строение, и поэтому ее целесообразно изучать по частям — выделяя более простые функции, содержательно соответствующие разным уровням естественного языка, например, функции, преобразующие смысл в синтаксическую структуру **); функции, преобразующие синтаксическую структуру в линейную последовательность слов; функции, преобразующие структурную характеристику словоформ в реальную словоформу, и т. п.***). Формальные грамма-

---

*) Ср., впрочем, замечания на стр. 150 и сл.

**) Здесь и ниже имеются в виду «заполненные» структуры, например, синтаксические деревья, в узлах которых помещены конкретные лексемы.

***) Как можно видеть из предшествующего изложения, авторы полагают, что более адекватной моделью языка был бы не механизм, производящий (перечисляющий) его правильные фразы, а устройство, преобразующее любой заданный смысл в соответствующие тексты или, наоборот, извлекающее смысл из любого заданного текста (ср. Жолковский—Мельчук 1967). Таким образом, наш подход отличается от концепции Н. Хомского. Было бы весьма интересно разобраться в данном отличии по существу; однако мы этого делать не будем, поскольку соответствующая проблема выходит за рамки задач этой книги, а способ ее решения практически не сказывается на изложении нашего основного материала, в частности теории формальных грамматик.

тики — это некоторый способ изучать и описывать подобные функции-компоненты. А именно, грамматики позволяют задавать множества значений этих функций. В отличие от самой функции, которая от любого указанного значения аргумента позволяет перейти к вполне определенному результату, существующие формальные грамматики описывают только совокупность возможных результатов, не давая прямых указаний, как именно можно получить результат, соответствующий определенному исходному «запросу».

Эту особенность рассматриваемые в книге формальные грамматики разделяют с обычными грамматиками (т. е. с грамматиками в общеупотребительном смысле слова).

Однако между обычными и формальными грамматиками имеется существенное различие. В формальных грамматиках все утверждения формулируются исключительно в терминах небольшого числа четко определенных и весьма элементарных «вещей» (символов и операций). Это делает формальные грамматики очень простыми с точки зрения их логического строения *) и облегчает изучение их свойств дедуктивными методами. Однако эта же самая особенность приводит к тому, что формальные грамматики оказываются весьма громоздкими: если мы хотим, чтобы разных типов исходных деталей было как можно меньше, а сами эти детали были как можно проще, то для описания достаточно сложных явлений естественного языка таких деталей, точнее, экземпляров таких деталей, требуется очень много. Поэтому формальные грамматики не способствуют повышению обозримости лингвистических описаний и неудобны для непосредственного использования человеком

*) Слово «простой» не следует понимать здесь в обиходном смысле как «наглядный», «легко обозримый» и т. п. Имеется в виду простота (= несоставность, элементарность) компонентов и простота (= единообразность) способа соединения компонентов в целое—так сказать, локальная простота. Само же целое может быть очень большим по объему и весьма сложным по строению, а потому плохо обозримым, т. е. вовсе не простым в буквальном смысле этого слова.

(например, учить иностранный язык с помощью его формальной грамматики вряд ли целесообразно). Формальные грамматики предназначаются сугубо для научного, теоретического исследования наиболее общих свойств языка. Впрочем, не исключается и их практическое использование, особенно в связи с применением вычислительных машин, например, при автоматическом переводе.

Мы не будем излагать теорию грамматик в полном объеме, а ограничимся основными сведениями о порождающих грамматиках (§§ 2—5); кроме того, в § 6 кратко характеризуются некоторые понятия, относящиеся к другим разделам теории грамматик (категориальные грамматики и автоматы с магазинной памятью).

Со следующего раздела начинается изложение конкретного материала — теории порождающих грамматик.

Предварительный пример: правила образования русских причастий

Допустим, что нас интересует совокупность всех форм причастий в письменной разновидности русского языка и мы хотим как-то задать эту совокупность, например, выписав систему правил, с помощью которых можно было бы получать любые правильные формы причастий, не получив при этом ни одной неправильной. (Подчеркнем, что данная задача является частным случаем общей задачи, стоящей перед формальными грамматиками: их цель состоит в задании совокупностей цепочек.)

Ниже приводится образец подобных правил, но не для всех русских глаголов, а только для нескольких, выбранных довольно случайным образом; при этом ради простоты примера правила составлены так, как если бы глаголов других типов и с другими особенностями не было вовсе. Тем самым наш пример не претендует на какую-либо значимость для описания русского языка и носит чисто иллюстративный характер.

Мы будем представлять себе словоформу-причастие как цепочку морф, включающую от трех до пяти морф: *вед* + *ш* + *ий*, *раздел'* + *а* + *jущ* + *ий*, *раздел'* + *а* + *jущ* + *ий* + *ся* *).

Словоформы и морфы записываются в основном в принятой орфографии с небольшой примесью транскрипции — там, где это удобно для формулировки правил: в ряде случаев особо обозначается мягкость согласных (только парных мягких!) и используется *j*.

Различается шесть классов морф:

1) основы — *раздел'*-, *строj*-, *потер'*-, *люб'*- и т. д.

2) «тематический элемент» («расширитель основы») — -*и*-/-*а*-/-*ова*-/-*у*-;

3) имперфективирующий суффикс (суффикс, служащий для образования форм несовершенного вида) — -*ива*-/-*ыва*-/-*а*-;

4) суффиксы причастий — -*ащ*-/-*ущ*-, -*вш*-/-*ш*- и т. д.;

5) окончания (флексии) причастий — -*ая*, -*ую*, -*ий*/-*ый* и т. д.

6) возвратная частица — -*ся*.

В правилах, по которым морфы соединяются друг с другом, существенным образом используются определенные признаки ряда морф **), а именно:

---

*) В рамках данного примера корень с префиксом рассматриваются вместе как одна морфа («основа»). Апостроф обозначает мягкость согласных, т. е. *л'а* = *ля*, *н'о* = *нё* и т. п.

**) Пояснения к обозначениям значений признаков морф: «сов-несов» означает, что данный глагол омонимичен в отношении вида (*автоматизировать*, *исследовать*);

«а» означает необходимость тематического элемента -*а*- (*опозд* + *а*-), «ова» — необходимость тематического элемента -*ова*-/-*у*-(*исслед* + *ова*-, *исслед* + *у* + *ют*); «ã» означает, что тематический элемент -*а*- возможен, но не необходим (*стон* + + *а* + *вш* + *ий*, но *стон* + *ущ* + *ий*), «ĩ» — что тематический элемент -*и*- возможен, но не необходим (*крас* + *и* + *вш* + + *ий*, но *крас* + *ящ* + *ий*); «атем» означает невозможность тематического элемента (*вед* + *ш* + *ий*);

«ЫВА» означает возможность присоединения имперфективирующего суффикса -*ыва*-/-*ива*- (*опаздывать*), «А» — имперфективирующего суффикса -*а*- (ср. *решить* — *решать*), «∅» — невозможность имперфективирующего суффикса (*построить*);

Для основ: 1) переходность/непереходность (t/i); 2) вид (сов/несов/сов-несов); 3) спряжение (I/II); 4) возможность или необходимость тематического элемента (а/ова/ã/и̃/атем); 5) возможность присоединения имперфективирующего суффикса (ЫВА/А/ϕ); 6) возможность или необходимость присоединения *-ся* (ся/$\overline{\text{ся}}$/ся-$\overline{\text{ся}}$).

Для суффиксов причастий: 1) спряжение (I/II/I-II); 2) залог (активное/пассивное = = акт/пасс); 3) время (настоящее/прошедшее = наст/прош).

Для флексий причастий — форма (полная/краткая = пф/кф).

Ниже приводятся перечни морф всех названных классов с указанием нужных признаков.

Класс 1. Основы

автоматизир-(t, сов-несов, I, ова, ϕ, ся-$\overline{\text{ся}}$)
вед-(t, несов, I, атем, ϕ, ся-$\overline{\text{ся}}$)
исслед- (t, сов-несов, I, ова, ϕ, ся-$\overline{\text{ся}}$)
крас'-(t, несов, II, и̃, ϕ, ся-$\overline{\text{ся}}$)
люб'- (t, несов, II, и̃, ϕ, ся-$\overline{\text{ся}}$)
нес-(t, несов, I, атем, ϕ, ся-$\overline{\text{ся}}$)
опозд-(i, сов, I, а, ЫВА, $\overline{\text{ся}}$)
покрас'-(t, сов, II, и̃, ϕ, ся-$\overline{\text{ся}}$)
построј-(t, сов, II, и̃, ϕ, ся-$\overline{\text{ся}}$)
потер'-(t, сов, I, а, ϕ, ся-$\overline{\text{ся}}$)
привед-(t, сов, I, атем, ϕ, $\overline{\text{ся}}$)
принес-(t, сов, I, атем, ϕ, $\overline{\text{ся}}$)
раздел'-(t, сов, II, и̃, А, ся-$\overline{\text{ся}}$)
смеј-(i, несов, I, ã, ϕ, ся)

«ся» означает необходимость присоединения *-ся* (*смеяться*), «ся» — невозможность присоединения *-ся* (*стонать*), «ся-ся» — возможность форм как с *-ся*, так и без *-ся* (*терять — теряться*);

«I-II» означает, что данный суффикс может присоединяться к основам как I, так и II спряжения (*терявший — красивший*).

спрос'-(*t*, сов, II, и̃,ЫВА,ся-$\overline{\text{ся}}$)
стон-(*i*, несов, I, ã, *ɸ*, $\overline{\text{ся}}$)
строj-(*t*, несов, II, и̃, *ɸ*, ся-$\overline{\text{ся}}$)
тер'-(*t*, несов, I, а, *ɸ*, ся-$\overline{\text{ся}}$)

Класс 2. Тематические элементы

-а-	*-ова-*
-и-	*-у-*

Класс 3. Имперфективирующие суффиксы

-а-
-ыва-

Класс 4. Суффиксы причастий

-ащ-(II, акт, наст)	*-им*-(II, пасс, наст)
-ущ-(I, акт, наст)	*-ом*-(I, пасс, наст)
-вш-(I-II, акт, прош)	*-нн*-(I, пасс, прош)
-ш- (I-II, акт, прош)	*-онн*-(II, пасс, прош)

Класс 5. Флексии причастий

-ая	*-ую*	(пф)	-Λ	(кф)
-ого	*-ые*		*-а*	
-ое	*-ый*		*-о*	
-ой	*-ым*		*-ы*	
-ом	*-ыми*			
-ому	*-ых*			

(Λ—«нулевое окончание», как в *потерян*)

Класс 6. Возвратная частица

-ся

Теперь мы приведем правила, в соответствии с которыми из перечисленных морф строятся формы русских причастий.

I. Правила общего строения

1. В словоформу должно входить не более чем по одной морфе из каждого класса.

2. Морфы должны следовать друг за другом в порядке нумерации классов.

3. Морфы классов 1, 4, 5 (основа + суффикс причастия + флексия) должны присутствовать обязательно.

II. Правила несовместимости

Словоформа не может содержать одновременно:

1. Морфы классов 2 и 3 (тематический элемент и имперфективирующий суффикс) *).

2. Основу с признаком «*i*» и суффикс причастия с признаком «пасс» (от непереходных глаголов невозможны страдательные причастия).

3. Основу с признаком «$\overline{\text{ся}}$» и частицу *-ся*.

4. Основу с признаком «сов» при отсутствии имперфективирующего суффикса и суффикс причастия с признаком «наст» (от глаголов сов. вида невозможны причастия наст. времени).

5. Основу с признаком «*ф*» и имперфективирующий суффикс.

6. Основу с признаком «I» и без признака «атем» и суффикс причастия с признаком «II» (тематические глаголы I спряжения не допускают суффиксов II спряжения).

7. Основу с признаком «атем» и суффикс причастия с признаком «II», отличный от *-онн-* (атематические глаголы не допускают суффиксов II спряжения, за исключением *-онн-*).

8. Имперфективирующий суффикс и суффикс причастия с признаком «II» (имперфективирующий суффикс переводит любой глагол в первое спряжение).

---

*) В целях упрощения примера данное правило не учитывает многочисленных форм типа *арестовывавший*, *выковывающий*, *организовывавший* и т. п.

9. Основу с признаком «II» при отсутствии имперфективирующего суффикса и суффикс причастия с признаком I.

10. Суффикс причастия с признаком «акт» и флексию с признаком «кф» (действительные причастия не имеют краткой формы *)).

11. Основу с признаком «атем» (соответственно без признака «атем») и суффикс *-вш-* (соответственно *-ш-*); ср. *ведший*, но *опоздавший*.

12. Основу с признаком «й̃» или «ā», тематический элемент и суффикс причастия, начинающийся на гласный (если при данной основе тематический элемент не обязателен, то перед суффиксом причастия, начинающимся на гласный, он не используется).

13. Основу с признаком «й̃» (соответственно с признаком «ā») и тематический элемент, отличный от *-и-* (соответственно от *-а-*).

14. Имперфективирующий суффикс и суффикс причастия с признаками «пасс», «прош», ср. **спраш-ива-нн-ый*, **раздел'-а-нн-ый*.

15. Суффикс причастия с признаком «пасс» и частицу *-ся* (страдательные причастия не могут быть возвратными).

III. Правила неотделимости

Словоформа обязательно должна содержать:

1. При наличии основы с признаком «а» — либо тематический элемент *-а-*, либо имперфективирующий суффикс.

2. При наличии основы с признаком «ова» — либо тематический элемент *-ова-* (если имеется суффикс причастия с признаком «прош»), либо тематический элемент *-у-* (если имеется суффикс причастия с признаком «наст»).

*) Известное огрубление: в поэтической речи краткие действительные причастия изредка встречаются. Ср., например, ... *их воздухом поющ тростник и сладок* (О. Мандельштам).

3. При наличии основы, не имеющей признака «атем», и суффикса причастия с началом на согласный — либо тематический элемент, либо имперфективирующий суффикс.

4. При наличии основы с признаком «ся» — частицу *-ся*.

IV. Морфонологические и фонологические *) правила

1. Между двумя соседними гласными, принадлежащими к разным морфам, появляется *j*. Здесь имеются в виду случаи типа *раздел'* + *а* + *ущ* + → → *раздел'* + *а* + *jущ* +.

2. В словоформе, содержащей суффикс *-ыва-*, корневой гласный *о* (последний *о* в основе) меняется на *а* **).

3. Перед суффиксами *-онн-*, *-ыва-* конечный согласный основы *-с'-* заменяется на *-ш-*, а конечный согласный *-б'-* — на *-бл'-* (аналогично, *д'* → *ж*, *т'* → *ч*, *в'* → *вл* и т. д.; но в нашем списке нет основ на *-д'-*, *-т'-*, *-в'-*).

4. Перед суффиксом *-онн-* твердые конечные согласные атематических основ смягчаются: *д* → *д'*, *с* → *с'* и т. д.; *принес* + *у* / *принес'* + *онн* + *ый*.

5. Перед флексией с признаком «кф» суффиксы *-онн-* и *-нн-* преобразуются соответственно в *-он-* и *-н-*.

6. Сочетание *jи* заменяется на *и*.

---

*) Морфонологические правила — это правила, относящиеся к последовательностям фонем (в нашем случае фонемы для простоты отождествляются с буквами), но обязательно учитывающие морфологическую роль этих последовательностей. Фонологические правила имеют дело просто с последовательностями фонем, безотносительно к их морфологическому статусу. В данной группе правил фонологическим является только правило IV.6.

**) Чтобы описать неучтенные здесь случаи типа *основ* + + *ать* — *основ* + *ывать* (чередование *о*/*а* невозможно) или *удосто* + *ить* — *удосто* + *ивать* — *удоста* + *ивать* (чередование *о*/*а* возможно, но не обязательно), было бы необходимо ввести еще один признак основ: чередование *о*/*а* перед *-ыва-* возможно/невозможно/обязательно.

V. Графическо-орфографические правила

1. Сочетания *ja*, *jy*, *jo* изображаются буквами *я*, *ю*, *е* *) соответственно.

2. Сочетания *X'a*, *X'y*, *X'o*, *X'u*, *X'ы* изображаются на письме как *Хя*, *Хю*, *Хе*, *Хи*, *Хи* соответственно (*X'* — любой парный мягкий согласный).

3. После букв *ж*, *ч*, *ш*, *щ* вместо *ы*, *о* пишутся *и*, *е*.

Примечание. Особую трудность представляет образование страдательных причастий прошедшего времени от основ несов. вида без имперфективирующего суффикса. В одних случаях они явно возможны — *писанный*, *крашенный*, в других, по-видимому, явно невозможны — **веденный*, **любленный*; имеется много промежуточных, не вполне ясных случаев: *терянный*? *строенный*? Как кажется, решающее значение здесь имеет узус, который в изложенных правилах не учитывается, а потому подобные образования, формально всегда возможные, допускаются этими правилами (и порождаются построенной на основе этих правил грамматикой, стр. 34—43).

Возможно, читателя удивит большое количество и разнообразие правил, оказавшихся необходимыми для описания причастий от нескольких русских глаголов. Следует признаться, что этот факт удивил и самих авторов, которые стремились подобрать пример, как можно более простой, но вместе с тем показательный. Однако с фактами приходится считаться: русская морфология заслуженно считается очень сложной. В предложенных правилах эта сложность гораздо более заметна, чем в «обычных» описаниях, по той причине, что здесь все выписано в явной форме, тогда как в существующих описательных грамматиках русского языка этого никогда не делается. Впрочем, надо иметь в виду, что нам пришлось включить в число правил и такие, которые совсем не

*) Это *е* (равно как и *е* в пп. 2—3) может соответствовать как *е*, так и *ё*, которые в печатном тексте обычно не различаются. Выбор между ними определяется ударением: под ударением *е* = *ё* (*принесенный*), без ударения *е* = *е* (*теряемый*).

являются специфическими для причастий (в частности, все графическо-орфографические правила). Поэтому — и это особенно существенно — если значительно увеличить наш список основ (т. е. привлечь намного больше глаголов), то это не потребовало бы сколько-нибудь значительного увеличения объема правил.

Определение и пример порождающей грамматики

Итак, мы построили точные правила, описывающие нужную нам совокупность словоформ — формы причастий нескольких русских глаголов—и тем самым решили задачу, сформулированную на стр. 26. Однако наше описание неудовлетворительно в одном важном отношении, а именно: оно само не построено по каким-либо определенным правилам, а образующие его правила не состоят из заранее выделенных и фиксированных элементарных компонентов. (Разумеется, такие компоненты можно выделить; однако для этого требуется специальное исследование, и мы получим совсем другое описание.) Если же мы хотим не только описывать конкретные языки, но и рассматривать в самом общем виде *способы описывать языки*, то необходимо строить все наши описания *единообразно*, составляя их из элементарных компонентов, соединяемых по строго определенным правилам. Только в этом случае мы сможем применять к лингвистическим описаниям строгие (математические) рассуждения.

Формальная грамматика и представляет собой такого рода описание. Весь арсенал используемых ею средств четко фиксируется в ее определении, а все ее утверждения имеют точно определенную форму, также фиксируемую определением.

Как уже было сказано, мы будем рассматривать только порождающие формальные грамматики. Перейдем к определению соответствующего понятия. Одновременно будем строить в качестве иллюстрации грамматику Γ_0, порождающую те же формы причастий, что и в разобранном выше примере.

Однако предварительно необходимо сделать два важных замечания.

Во-первых, на самом деле мы будем рассматривать не любые мыслимые порождающие грамматики, а определенный класс таковых, введенный Н. Хомским (Хомский 1962, 1965а). Именно с грамматик этого класса началось изучение порождающих грамматик вообще; более того, и сейчас грамматики Хомского остаются наиболее исследованными во всех аспектах: общая теория формальных порождающих грамматик на 90%, если не больше, сводится к теории грамматик Хомского. Это позволяет нам вместо полного названия «формальные порождающие грамматики в смысле Хомского» пользоваться термином «порождающие грамматики», или даже просто «грамматики» (разумеется, только в тех случаях, когда контекст исключает неправильное понимание).

Во-вторых, в работах по теории порождающих грамматик (мы имеем в виду прежде всего работы на русском языке, в том числе и переводные) имеет место явно нежелательный терминологический разнобой. Для ряда понятий употребляется по два-три разных термина, причем многие из этих терминов, и часто как раз самые употребительные, по своей внутренней форме представляются нам неудачными (обычно это объясняется слишком буквальным переводом, даже калькированием английских терминов). Поэтому мы позволим себе в отдельных случаях пользоваться малоупотребительными терминами, не приводя, однако, в тексте тех соображений, которые заставили нас предпочесть тот или иной термин: такие соображения, а также параллельные термины на русском и английском языках приведены в комментированном словарике терминов (см. Приложение II).

Порождающая грамматика — это система, состоящая из четырех частей: основной, или терминальный, словарь; вспомогательный словарь; начальный символ; набор правил подстановки.

1. Основной (терминальный) словарь — набор исходных элементов, из которых строятся цепочки, порождаемые грамматикой.

В грамматике Γ_0 — это набор всех русских букв, которые набираются курсивом в отличие от прочих употреблений этих букв (в тексте книги, в качестве фонетической транскрипции, для обозначения информаций к вспомогательным символам и т. д.). Таким образом, реальные русские морфы и словоформы будут записываться курсивом.

Элементы основного словаря называют **основными (терминальными) символами**.

Морфы как таковые не считаются самостоятельными символами и поэтому не входят ни в основной, ни во вспомогательный (см. ниже, п. 2) словарь: они рассматриваются как цепочки терминальных символов — букв. Роль словаря основ (ср. 1-й класс морф в только что разобранном примере, стр. 28—29), здесь по существу играют правила группы VIII (см. ниже, стр. 41).

Замечание. С лингвистической точки зрения было бы правильнее (и естественнее) описывать формы русских причастий с помощью **двух разных** грамматик: одна из них представляла бы причастие в виде последовательности символов морф, а вторая перерабатывала бы эти последовательности в реальные буквенные цепочки, т. е. строила бы формы русских причастий в правильной орфографической записи. Тем самым разные уровни языка описывались бы разными грамматиками. (Вторая грамматика, по сути дела, не вполне удовлетворяла бы определению порождающей грамматики *), хотя формально ее можно было бы сделать таковой.) Тогда основной словарь первой грамматики состоял бы из символов морф (перечень основ и аффиксов), а вспомогательный содержал бы только символы категорий. Вторая грамматика имела бы в основном словаре русские буквы, а во вспомогательном словаре — символы морф, транскрипционные знаки, граничный символ и, возможно, категории букв (гласные/согласные и т. п.). Подчеркнем, что вторая грам-

*) А именно, она порождала бы «выходные» цепочки, исходя не из единственного начального символа (см. стр. 38, п. 3), а из «выходных» цепочек первой грамматики.

матика ни по строению, ни по назначению не связана специально с порождением причастий: она представляет морфологический, фонологический и графическо-орфографический уровни языка и необходима для порождения любых словоформ (ср. замечание в конце предыдущего раздела, стр. 33—34). При таком разбиении основные и вспомогательные словари обеих грамматик выглядели бы более естественно. Мы, однако, предпочли — ради единства примера — строить одну грамматику.

2. Вспомогательный (нетерминальный) словарь — набор символов, которыми обозначаются классы исходных элементов или цепочек исходных элементов, а также, в отдельных случаях, некоторые специальные элементы. Эти символы называются в с п о м о г а т е л ь н ы м и, или н е т е р м и н а л ь н ы м и. В грамматике Γ_0 мы введем следующие нетерминальные символы:

ПРИЧ — причастие;

ПРИЧ (x, y, z) — причастие заданного залога, времени и возвратности (значения для x, y, z указаны ниже, правило I);

$O'(a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ — основа причастия, включая имперфективирующий суффикс или тематический элемент, если таковые есть, и вспомогательная информация к ней: a_1 — переходность (t/i); a_2 — вид (сов/несов); a_3 — спряжение (I/II); a_4 — тематичность (а/ова/$\widetilde{\text{а}}$/$\widetilde{\text{и}}$/атем); a_5 — возможность имперфективирующего суффикса (ЫВА/А/ϕ); a_6 — возможность возвратной формы (ся/$\overline{\text{ся}}$) *); a_7 — возможность невозвратной формы (¬ся/¬$\overline{\text{ся}}$);

$O(a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ — «чистая» основа причастия (без тематического элемента и без имперфек-

*) Здесь «ся» означает возможность возвратной формы, «$\overline{\text{ся}}$» — невозможность возвратной формы, «¬ ся» — возможность невозвратной формы, «¬ $\overline{\text{ся}}$» — невозможность невозвратной формы. Обратим внимание на то, что обозначения «ся» и «$\overline{\text{ся}}$» имеют теперь не тот смысл, что на стр. 28—32: там возвратность и невозвратность описывались одним признаком, принимающим три значения, а здесь — двумя бинарными признаками a_6 и a_7, что позволяет сократить число правил.

тивирующего суффикса) с той же вспомогательной информацией, что и выше;

С (x, y, a_3)— суффикс причастия с информацией к нему (x — залог, y — время, a_3 — спряжение, см. выше);

Ф (u) — флексия причастия с информацией к ней; u — форма (краткая/полная — кф/пф);

О, С, Ф — названные морфы без информаций;

И — имперфективирующий суффикс;

Т — тематический элемент;

+ — граница между морфами; автоматически появляется после тех морф, которыми не могут оканчиваться словоформы;

X' — мягкая согласная; здесь X — обозначение произвольной согласной;

j — обозначение звука [j] (йот).

3. **Начальный символ** — выделенный нетерминальный символ, обозначающий совокупность (класс) всех тех языковых объектов, для описания которых предназначается данная грамматика. В грамматике $Г_0$ — это символ «ПРИЧ», так как наша цель — описать совокупность причастий. (В грамматике, порождающей предложения, начальным будет символ, означающий «предложение»; в грамматике, порождающей допустимые слоги, начальный символ означает «слог» и т. д.)

4. **Правила подстановки** — выражения вида «$X \rightarrow Y$», что означает «заменить X на Y» или «подставить Y вместо X», где X и Y — цепочки, содержащие любые терминальные или нетерминальные символы. В грамматике $Г_0$ правила подстановки таковы:

I. З а д а н и е г р а м м а т и ч е с к и х з н а ч е н и й п о р о ж д а е м о г о п р и ч а с т и я*).

---

*) Для лучшей обозримости правила подстановки разбиты на группы (нумеруемые римскими цифрами), каждая из которых отвечает определенной содержательной задаче; эта задача указывается при номере группы. Номера групп и правил не следует понимать как указания о порядке их применения: порядок применения правил порождающей грамматики произволен, см. стр. 46.

ПРИЧ → ПРИЧ (x, , z).
Здесь x = акт, пасс;
y = наст, прош;
z = возвр, невозвр.
При x = пасс необходимо, чтобы z = невозвр.

Запись «ПРИЧ → ПРИЧ (x, y, z)» применяется для сокращения: на самом деле здесь написано не одно, а шесть правил, соответствующих допустимым наборам значений переменных x, y, z, например «ПРИЧ → ПРИЧ (акт, наст, возвр)», «ПРИЧ → ПРИЧ (пасс, наст, невозвр)» и т. д. Во всех остальных случаях переменные x, y, z используются таким же образом.

II. Реализация грамматических значений соответствующими морфами.

1. ПРИЧ (акт, наст, возвр) → → О′ (несов, a_3, ся) С (акт, наст, a_3) ***Фся****)
2. ПРИЧ (акт, прош, возвр) → → О′ (a_3, ся) С (акт, прош, a_3) ***Фся***
3. ПРИЧ (акт, наст, невозвр) → → О′ (несов, a_3, $\neg$ся) С (акт, наст, a_3) Ф
4. ПРИЧ (акт, прош, невозвр) → → О′ (a_3, $\neg$ся) С (акт, прош, a_3) Ф
5. ПРИЧ (пасс, наст, невозвр) → → О′ (t, несов, a_3, $\neg$ся) С (пасс, наст, a_3) Ф
6. ПРИЧ (пасс, прош, невозвр) → → О′ (t, a_3, $\neg$ся) С (пасс, прош, a_3) Ф

При записи информации к морфам для краткости опускаются обозначения тех признаков, которые в данном правиле могут принимать любые значения. Таким образом, например, запись «О(несов, ся)» есть сокращение для многих выражений вида «О (a_1 несов, a_3, a_4, a_5, ся, a_7), где a_1, a_3, a_4, a_5, a_7 принимают любые допустимые значения; соответственно запись «С (акт, наст)» — также сокращение для двух выражений вида «С (акт, наст, a_3)». Поэтому, например,

---

*) Как указывалось выше, курсивом набираются реальные русские морфы и словоформы.

строка II.1 фактически содержит в себе не одно правило, а 120 разных правил.

III. Разложение «сложной» основы (выделение собственно основы и тематического элемента или имперфективирующего суффикса, если таковые возможны).

1. O′(не атем) → O (не атем) T*)

2. O′(несов, не ϕ) C (x, y) → O (сов, не ϕ) ИС (x, y, I)

В этом правиле переменные x и y должны удовлетворять следующему условию: при x = пасс необходимо, чтобы y = наст.

3. O′(атем) → O(атем)

IV. Реализация тематического элемента соответствующей морфой.

1. O (ã) Tα → O (ã) α
2. O (ĩ) Tα → O (ĩ) α
3. O (a) T → O (a)*a* +
4. O (ова)TC (наст) → O(ова)*у* + C (наст)
5. O (ова) TC (прош) → O(ова)*ова* + C (прош)
6. O (ã, I)Tβ → O(ã, I)*a* + β
7. O (ĩ, II) Tβ → O(ĩ, II)*и* + β

Здесь α и β — сокращения: α — произвольная гласная, β — произвольная согласная.

V. Реализация имперфективирующего суффикса соответствующей морфой.

1. O (A) И → O (A) *a* +
2. O (ЫВА)И → O (ЫВА)*ыва* +

VI. Реализация суффикса причастия соответствующей морфой.

1. C (акт, наст, I) → *ущ* +
2. C (акт, наст, II) → *ащ* +
3. C (пасс, наст, I) → *ом* +
4. C (пасс, наст, II) → *им* +
5. O (не атем) XC (акт, прош) → O (не атем) *Хвш* +
6. O(атем) C(акт, прош) → O (атем)*ш* +

*) Здесь «не атем» употребляется для сокращения и означает любое значение признака a_4, отличное от «атем», т. е. «а», «ова», «ã» или «ĩ». Аналогичный смысл имеет запись «не ∅».

7. О(не атем) XС(пасс, прош, I) →
→О(не атем) *Хнн* +
8. О (атем) С (пасс, прош) → О (атем) *онн* +
9. С (пасс, прош, II) → *онн* +

Здесь X — сокращение: любой имперфективирующий суффикс или тематический элемент, например *у* +, *ыва* +.

VII. Выбор формы причастия (краткой или полной) и реализация флексии соответствующей морфой.

1. Ф → Ф (пф)
2. С (пасс)Ф → С (пасс)Ф (кф)

3—15. Ф (пф) → *ый*, *ое*, *ого*, *ому*, *ым*, *ом*, *ая*, *ой*, *ую*, *ые*, *ых*, *ым*, *ыми*

16—19. Ф (кф) → Λ, *о*, *а*, *ы*

Здесь Λ — пустая цепочка, т. е. цепочка, не содержащая никаких символов. Правило Ф(кф) → Λ содержательно означает «можно зачеркнуть символ Ф(кф)».

VIII. Реализация основы соответствующей морфой*).

О (t, несов, I, атем, ϕ, ся, $\neg$ся) → *вед* +, *нес* +, ...
О (t, несов, I, ова, ϕ, ся, $\neg$ся) → *исслед* +, *автоматизир* +, ...
О (t, несов, I, а, ϕ, ся, $\neg$ся) → *тер'* +, ...
О (t, сов, I, атем, ϕ, $\overline{\text{ся}}$, $\neg$ся) → *привед*+, *принес*+,...
О (t, сов, I, ова, ϕ, ся, $\neg$ ся) → *исслед* +, *автоматизир* +, ...
О (t, сов, I, а, ϕ, ся, $\neg$ ся) → *потер'* +, ...
О (t, несов, II, $\tilde{\text{и}}$, ϕ, ся, $\neg$ ся) → *строј* +, *крас'* +, *люб'* +, ...
О (t, сов, II, $\tilde{\text{и}}$, А, ся, $\neg$ ся) → *раздел'* +, ...
О (t, сов, II, $\tilde{\text{и}}$, ЫВА, ся, $\neg$ ся) → *спрос'* +, ...
О (i, несов, I, $\tilde{\text{а}}$, ϕ, $\overline{\text{ся}}$, $\neg$ся) → *стон* +, ...
О (i, несов, I, $\tilde{\text{а}}$, ϕ, ся, $\overline{\neg\text{ся}}$) → *смеј* +, ...
О (i, сов, I, а, ЫВА, $\overline{\text{ся}}$, $\neg$ся) → *опозд* +, ...

---

*) Правила группы VIII не нумеруются, поскольку каждая строчка здесь представляет целый набор правил, число которых определяется количеством привлекаемых основ с данной информацией.

IX. Морфонологические правила.

1. $\alpha_1 + \alpha_2 \to \alpha_1 + j\alpha_2$ (где α_1 и α_2 — произвольные гласные)

2. j + *и* → *и*

3. оX + *ыва* → аX + *ыва*

Здесь X — сокращенное обозначение произвольной цепочки не длиннее трех символов. Имеется в виду чередование *о/а* в глагольных корнях типа *опозд-ать/опазд-ывать*. Как кажется, группа согласных, следующая в корне за чередующимся *о* (т. е. отделяющая его от суффикса *-ыва-*), не может содержать больше трех букв.

4. с' + X → *ш* + X, б' + X → *бл'* + X — здесь X = *онн*, *ыва*

. . . .

5. *д* + *онн* → *д'* + *онн*
   с + *онн* → *с'* + *онн*

. . . .

6. *нн* + Ф (кф) → *н* + Ф(кф).

X. Графическо-орфографические правила.

1. *j*+*а* → + *я*
   j+*у* → + *ю*
   j+*о* → + *е*
   . . .
   jа→*я*
   jу→*ю*
   jо→*е*
   . . . .

2. X'+ *а* → X +*я*
   X'+ *у* → X +*ю*
   X'+ *о* → X +*е*
   X'+ *ы* → X +*и*
   X'+ *и* → X +*и*
   здесь X — любая согласная

3. X+*ы* → X+*и*, X+*о* → X+*е**) — здесь X = *ж*, *ч*, *ш*, *щ*

XI. Стирание показателя границы между морфами

$$X + Y \to XY$$

Здесь X и Y — любые такие морфы, что к $X + Y$ неприменимо ни одно из правил групп IX—X.

---

*) Данное правило безоговорочно верно лишь для *о* в глагольных суффиксах (т. е. в суффиксах, присоединяемых к глагольным основам).

Данное ограничение на X и Y препятствует тому, чтобы граница между морфами была стерта слишком рано, до применения соответствующих морфонологических правил. В противном случае такие морфонологические правила не были бы применены, а это привело бы к неправильным результатам (например, можно было бы получить **опоздывавший*, **спросенный*): дело в том, что если какое-либо морфонологическое правило м о ж е т быть применено, то оно д о л ж н о б ы т ь применено обязательно.

Понятия выводимости и вывода; язык, порождаемый грамматикой

Итак, на примере Γ_0 мы описали устройство формальной порождающей грамматики. Теперь введем следующие три понятия, необходимые для описания того, как применяется грамматика, т. е. для описания самого процесса порождения.

Н е п о с р е д с т в е н н а я в ы в о д и м о с т ь. Если имеются две цепочки X и Y, причем $X = Z_1AZ_2$, а $Y = Z_1BZ_2$ (Z_1 и/или Z_2 могут быть пустыми) и в грамматике Γ имеется правило $A \to B$, то Y непосредственно выводима в Γ из X. Другими словами, X может быть переработана в Y за один шаг — применением одной подстановки: Y получается из X подстановкой B на место некоторого вхождения цепочки A. Например, из цепочки

«О(t,сов,II,$\tilde{й}$,А,ся,$\neg$ся) *и*+С (акт, прош, II) Ф»

по правилу VI.5 непосредственно выводима цепочка

«О (t, сов, II, $\tilde{й}$,А, ся, $\neg$ся) *и* + *вш*+Ф».

В ы в о д и м о с т ь. Если имеется последовательность цепочек X_0, X_1, ..., X_n, в которой каждая следующая цепочка непосредственно выводима из предыдущей, то X_n в ы в о д и м а из X_0; сама же последовательность X_0, X_1, ..., X_n называется в ы в о д о м X_n из X_0. Это означает, что X_0 перерабатывается в X_n не обязательно за один шаг,

а последовательным применением нескольких подстановок. Очевидно, что непосредственная выводимость есть частный случай выводимости.

Приведем п р и м е р в ы в о д а в изложенной грамматике Γ_0 *).

ПРИЧ

(I) ПРИЧ (пасс, прош, невозвр)
(II.6) O′ (t, сов, II, и̃, ϕ, ся, ⌉ся) С (пасс, прош, II) Ф
(III.1) O (t, сов, II, и̃, ϕ, ся, ⌉ся) ТС (пасс, прош, II) Ф
(VI.9) O (t, сов, II, и̃, ϕ, ся, ⌉ся) Т*онн* + Ф
((IV.2) O (t, сов, II, и̃, $\tilde{\phi}$, ся, ⌉ся) *онн* + Ф
(VII.1) O (t, сов, II, и̃, ϕ, ся, ⌉ся) *онн* + Ф (пф)
(VII.15) O (t, сов, II, и̃, ϕ, ся, ⌉ся) *онн* + *ыми*
(VIII . .′ .) *покрас’* + *онн* + *ыми*
(IX.4) *покраш* + *онн* + *ыми*
(X.3) *покраш*+*енн*+*ыми*
(XI) *покрашенными*

Такой вывод (начинающийся начальным символом и заканчивающийся цепочкой, состоящей только из терминальных символов) называется п о л н ы м. Разумеется, не всякий вывод, начинающийся начальным символом, является полным; возможны, в частности, и такие выводы, начинающиеся начальным символом, которые невозможно продолжить до полного вывода,— «т у п и к о в ы е выводы».

Приведем п р и м е р т у п и к о в о г о в ы в о д а в Γ_0.

ПРИЧ

(I) ПРИЧ (акт, наст, возвр)
(II.1) O′(t, несов, II, атем, ϕ, ся, ⌉ся) С (акт, наст, II) Ф*ся*
(III.3) O(t, несов, II, атем, ϕ, ся, ⌉ся) С(акт, наст, II) Ф*ся*

---

*) В скобках (слева от строки вывода) указывается номер правила, применение которого к предыдущей строке дает данную.

(VI.2)O(t, несов, II, атем, ϕ, ся, $\neg$ся) *ащ* + Фся
(VII. 1) O(t, несов, II, атем, ϕ, ся, $\neg$ся) *ащ* +
+ Ф(пф)*ся*
(VII.11)O (t, несов, II, атем, ϕ, ся, $\neg$ся) *ащ* + *уюся*
(XI) O(t, несов, II, атем, ϕ, ся $\neg$ся)*ащуюся*

Этот вывод продолжить невозможно, хотя он и не является полным, т. е. не кончается цепочкой терминальных символов (русских букв). В грамматике Γ_0 нет правила, левая часть которого содержалась бы в последней цепочке данного вывода. Это объясняется тем, что в русском языке не существует атематических глаголов, спрягающихся по II спряжению. Таким образом, признаки «атем» и «II», которые при построении грамматики Γ_0 трактовались как независимые, в действительности являются связанными. Их связанность можно было бы учесть, однако это повело бы к заметному усложнению грамматики. Мы предпочли не поступать так, поскольку наличие в грамматике тупиковых выводов не является, вообще говоря, ее недостатком. От «хорошей» грамматики вовсе не требуется, чтобы любой вывод в ней заканчивался правильной терминальной цепочкой: достаточно, чтобы любой **полный** вывод давал правильную цепочку (в нашем случае — форму русского причастия).



Подчеркнем, что порождающая грамматика **не** является алгоритмом *): правила подстановки — это не последовательность предписаний, а совокупность разрешений. Это означает, что, во-первых, правило вида $A \to B$ понимается в грамматике как «A можно заменить на B» (а можно и не заменять), тогда как в алгоритме $A \to B$ означало бы «A следует заменить на B» (нельзя не заменить); во-вторых, порядок применения правил в грамматике произво-

*) Разъяснять здесь, что такое алгоритм, мы не можем, однако для понимания данной книги знать это необходимо. Читателю, не уверенному, что он в достаточной степени знаком с понятием алгоритма, рекомендуется обратиться хотя бы к статье «Алгоритм» в «Философской энциклопедии», т. 1 (М., 1960), 38—42, или к книге Б. А. Трахтенброта (1960).

лен (любое правило разрешается применять после какого угодно), тогда как в алгоритме был бы задан жесткий порядок применения отдельных инструкций. Однако необходимо помнить, что одно дело — разрешение применить правило, а другое дело — возможность сделать это: правило удается применить лишь к такой цепочке, которая содержит вхождение его левой части (так, правило $AB \to CD$ можно применить к цепочке $AAEABC$, но не к цепочке $AAAEBC$). Поэтому фактически порядок применения правил может диктоваться ими самими. Например, в грамматике Γ_0 правила могут применяться в процессе вывода только в некотором определенном, почти жестком порядке, хотя для Γ_0, как и вообще для любой грамматики, никаких внешних ограничений на порядок работы правил не существует. Так, в любом полном выводе никакое правило группы III не может быть применено раньше правила группы II, поскольку в левые части правил группы III входит символ O′, который может появиться в обрабатываемой цепочке только в результате применения одного из правил группы II; аналогично, правило IV.1 или IV.6 не удастся применить раньше правила группы VI и т. п.

Чтобы лучше освоиться с понятиями грамматики и вывода, читателю будет полезно построить несколько полных выводов в грамматике Γ_0, т. е., действуя чисто механически, породить ряд словоформ причастий. Для этого надо взять начальный символ «ПРИЧ», выбрать любое правило подстановки, имеющее в левой части этот символ, и применить его; далее, выбрать любое правило, применимое к полученному результату, применить его и т. д. В качестве образца можно пользоваться примером вывода на стр. 44.

Я з ы к, п о р о ж д а е м ы й г р а м м а т и к о й. Совокупность всех терминальных цепочек *), выводимых из начального символа в грамматике Γ,

*) Терминальная цепочка — цепочка, состоящая только из терминальных символов, см. стр. 36.

называется **языком, порождаемым грамматикой**, и обозначается L (Г). Следует подчеркнуть, что такое употребление термина «язык», введенное Н. Хомским, не совпадает с более принятым в лингвистике (а также и в математике) его толкованием. Хомский, различая систему и бесконечное, вообще говоря, множество результатов ее работы, называет первую грамматикой, а второе — языком. Обычно же в лингвистике, начиная с Ф. де Соссюра, языком называют, как известно, именно систему, а результат ее работы — речью. Однако, говоря о порождающих грамматиках, пользоваться такой терминологией по ряду причин (в основном языкового характера) неудобно; ср., например, **различные речи, порождаемые языками* и т. п. К тому же терминология Хомского стала общепринятой и привычной в данной области. Поэтому и мы будем пользоваться ею при изложении теории порождающих грамматик. В других случаях термин «язык» используется нами в обычном для лингвистики смысле; однако из контекста всегда видно, какой смысл имеется в виду.

В нашем примере язык, порождаемый грамматикой $Г_0$,— это совокупность всех причастий от указанных выше глаголов. Очевидно, что этот язык конечен. Однако грамматики могут порождать и бесконечные языки — ср. хотя бы грамматику $Г_1$, стр. 55—57.

Таким образом, применение грамматики — это построение полных выводов; последние цепочки этих выводов и образуют язык, порождаемый грамматикой.

Отметим, что возможна ситуация, когда две разные грамматики порождают один и тот же язык, т. е. одно и то же множество терминальных цепочек; в этом случае грамматики называются **эквивалентными**.

Теперь, подводя итог, мы можем окончательно сформулировать определение формальной порождающей грамматики (в дальнейшем мы будем говорить для краткости просто «грамматика»):

Грамматика есть упорядоченная четверка $\langle \boldsymbol{V}, \boldsymbol{V}_1, \boldsymbol{I}, \boldsymbol{S} \rangle$, где 1)—2) $\boldsymbol{V}$ и $\boldsymbol{V}_1$—не пересекающиеся конечные множества символов, называемые соответственно основным и вспомогательным словарями, 3) $\boldsymbol{I}$— выделенный элемент из $\boldsymbol{V}_1$, называемый начальным символом, 4) $\boldsymbol{S}$ — конечное множество выражений вида $A \to B$, где A и B — цепочки, состоящие из основных и вспомогательных символов (**цепочки над** $\boldsymbol{V} \cup \boldsymbol{V}_1$), а «$\to$»—символ, не принадлежащий ни $\boldsymbol{V}$, ни $\boldsymbol{V}_1$; эти выражения называются **правилами грамматики**, а множество $\boldsymbol{S}$ — **схемой грамматики**.

§ 3. Классы порождающих грамматик

Грамматики, отвечающие только что приведенному определению, представляют собой порождающие устройства очень общего характера: они способны $T_{1.1.1}$ порождать л ю б ы е (!) множества цепочек, какие только вообще могут быть порождены каким-либо автоматическим устройством *). Однако множество ЛВ фраз естественного языка — это множество, обладающее рядом специфических свойств. (Важное замечание: изучая фразы естественного языка в аспекте теории формальных грамматик, их обычно рассматривают как цепочки словоформ или морф, выступающих в роли терминальных символов.) В частности, естественно полагать, что для множества фраз существует распознающий алгоритм — способ узнавать относительно каждой предъявленной цепочки, является ли она фразой данного языка; при этом не обязательно предполагается деление всех фраз только на правильные и неправильные — можно допустить класс или классы промежуточных случаев. Это мнение вытекает из наблюдения фактов: носитель языка всегда может оценить правильность предложенной ему фразы, даже если он никогда не слышал ее раньше; стало быть, он пользуется каким-то распознающим алгоритмом, строение которого ему, однако, не известно. Более того, этот алгоритм должен обладать определенными специфическими свойствами: в частности, он всегда выдает ответ достаточно быстро.

*) Множества, порождаемые автоматическими устройствами произвольного вида, в теории алгоритмов называются рекурсивно перечислимыми. Тот факт, что любое рекурсивно перечислимое множество порождается некоторой грамматикой, строго доказан (Davies 1958, глава 6, § 2; Гладкий 1966, стр. 143; см. также идею доказательства в Хомский 1966, стр. 139).

Множества, для которых существуют распознающие алгоритмы, называются рекурсивными *), а те из них, для которых эти алгоритмы достаточно просты, т. е. выполняются в «не очень большое» число шагов, образуют еще более узкий класс и называются ниже «легко распознаваемыми множествами». Если мы хотим, чтобы грамматика порождала фразы естественного языка и только их, то на грамматики, определенные на стр. 48, необходимо наложить такие ограничения, чтобы порождаемые ими множества были рекурсивными и, более того, легко распознаваемыми. Представляется естественным, чтобы эти ограничения относились к правилам, т. е. к характеру замен: что именно и на что разрешается заменять.

Неукорачивающие грамматики

Можно начать с требования, чтобы в правиле вида $A \to B$ цепочка B не была короче цепочки A; тогда в процессе вывода цепочки не будут укорачиваться **). Оказывается, что уже этого, столь незначительного на первый взгляд, ограничения достаточно для нашей цели: языки, порождаемые «неукорачивающими» грамматиками, являются легко распознаваемыми! Этот важный факт строго доказывается, $\mathrm{T}_{1.2.13}$ причем доказательство, при всем его значении для математической лингвистики, весьма просто, и мы приведем его здесь, чтобы в какой-то степени ознакомить читателя с «кухней» математической лингвистики.

Рассмотрим неукорачивающую грамматику Г, содержащую p символов (терминальных и нетерми-

---

*) Всякое рекурсивное множество является рекурсивно перечислимым, обратное же неверно: не для всякого рекурсивно перечислимого множества существует распознающий алгоритм.

**) Под длиной цепочки понимается число символов в ней: например, длина цепочки $AABC+\ D$ равна 5. Заметим, что грамматика Γ_0 не удовлетворяет указанному требованию — в ней имеются «укорачивающие» правила, а именно IV.1, IV.2, IX.2, IX.6, X.1 и XI (см. стр. 40—42). Однако язык $L(\Gamma_0)$ является рекурсивным, хотя бы уже потому, что он конечен.

нальных вместе). Возьмем произвольную цепочку x длины n, состоящую из терминальных символов грамматики Г. Чтобы доказать, что L (Г) — легко распознаваемое множество, достаточно указать алгоритм, который для каждой такой цепочки за «не очень большое» число шагов решал бы, выводима она из начального символа грамматики Г или нет. Такой алгоритм существует; его идея заключается в следующем. Он строит один за другим всевозможные выводы в грамматике Г из начального символа $\boldsymbol{I}$. Построение каждого вывода состоит в том, что сначала к $\boldsymbol{I}$ применяется произвольное (применимое к $\boldsymbol{I}$) правило грамматики Г, затем к результату снова применяется какое-либо правило и т. д.; количество применений правил в данном выводе называется его длиной (заметим, что длина вывода равна числу цепочек в нем, не считая начального символа). Завершив построение каждого очередного вывода, алгоритм проверяет, не оканчивается ли он рассматриваемой цепочкой x. Если да, то процесс обрывается, так как тем самым получен ответ: цепочка x выводима из $\boldsymbol{I}$. Если нет, то процесс продолжается. Чтобы этот процесс не был бесконечным, алгоритм должен «знать», где остановиться, т. е. когда прекратить построение новых выводов. Этого можно добиться, если задать алгоритму такое конечное множество выводов M, что если ни один из выводов, входящих в M, не оканчивается цепочкой x, то и вообще никакой вывод не оканчивается этой цепочкой. Тогда, перебрав все выводы из M и не найдя среди них ни одного, оканчивающегося на x, алгоритм должен прекратить работу и выдать отрицательный ответ: x не выводима из $\boldsymbol{I}$. При этом еще недостаточно, чтобы множество M просто существовало; необходимо также, чтобы мы умели эффективно находить M по заданной цепочке x.

Укажем способ построения множества M по произвольной цепочке x. Если цепочка x длины n выводима из начального символа $\boldsymbol{I}$ грамматики Г, то имеется вывод — последовательность цепочек, ведущая от $\boldsymbol{I}$ к x. Можно считать, что в этом выводе ни

одна цепочка не повторяется: в противном случае образовавшуюся «петлю» можно безболезненно выкинуть. Поскольку наша грамматика неукорачивающая, ни одна из цепочек этой последовательности не может быть длиннее цепочки x, т. е. иметь длину больше n. (Именно здесь и использован существенным образом тот факт, что грамматика Г является неукорачивающей!) Число же различных цепочек длины. не превосходящей n, которые можно составить из p символов, заведомо не больше чем

$$p^n + p^{n-1} + p^{n-2} + \ldots + p^2 + p^1 + p^0 = \frac{p^{n+1}-1}{p-1} \text{ *)}$$

(сумма геометрической прогрессии) $< p^{n+1}$.

Стало быть, вывод цепочки x из I, если он существует, обязательно содержится среди всех возможных бесповторных последовательностей, составленных из $p^{n+1} = P$ разных цепочек. А таких последовательностей заведомо не больше чем

$$P! + C_P^1 \cdot (P-1)! + C_P^2 \cdot (P-2)! + \ldots$$
$$\ldots + C_P^{P-2} \cdot 2! + C_P^{P-1} \cdot 1! \text{ **)}$$

---

*) Число цепочек длины n из p символов равно p^n, число цепочек длины $(n-1)$ из p символов равно p^{n-1} и т. д.; число цепочек из 0 символов равно $p^0 = 1$ (цепочка длины 0 — пустая цепочка — единственна). Заметим еще, что здесь всегда $p > 1$, поскольку основной и вспомогательный словари предполагаются непустыми.

**) Поясним, как получается это выражение. Если имеется P разных элементов и мы строим из них бесповторные последовательности, то длина каждой такой последовательности не превосходит P: она равна P, если в последовательность входят все элементы, или меньше P в противном случае. Последовательностей длины P из P элементов имеется $P!$ (число перестановок из P элементов). Последовательностей длины $P-1$ из данных P элементов имеется $(P-1)! \cdot P$ (из данных $P-1$ элемента можно составить $(P-1)!$ последовательностей, а выбрать $P-1$ элемент из P элементов можно $P = C_P^1$ способами). Аналогично, последовательностей длины $P-2$ из данных P элементов имеется $(P-2)! \cdot C_P^2$ (поскольку $P-2$ элемента можно выбрать из данных P элементов именно C_P^2 способами) и т. д.

В этой сумме P членов; ее k-й член равен

$$C_P^k \cdot (P-k)! = \frac{P!}{k!\,(P-k)!} \cdot (P-k)! = \frac{P!}{k!} \leqslant P!.^{*)}$$

Поэтому вся сумма не больше чем $P! \cdot P < (P+1)! = (p^{n+1}+1)! < (p^{n+2})!$. Те из полученных $(p^{n+2})!$ последовательностей, которые являются выводами в грамматике Г, и образуют искомое множество M. Перебирая их, мы либо обнаружим нужный вывод, либо убедимся, что его нет.

Указав алгоритм распознавания выводимости произвольной цепочки в грамматике Г, мы доказали тем самым, что L (Г) — рекурсивное множество; поскольку этот алгоритм дает ответ не более чем в $(p^{n+2})!$ шагов **) для любой цепочки длины n, то L (Г) — легко распознаваемое множество. Следует отметить, что понятие «не очень большое» (число шагов) — вещь условная; на самом деле, число $(p^{n+2})!$, очевидно, настолько велико, что предложенный алгоритм практически неосуществим даже с помощью самых мощных ЭЦВМ. Однако в теории алгоритмов часто встречаются рекурсивные множества, для распознавания которых требуется несравненно большее число шагов, и даже такие рекурсивные множества, для которых число шагов распознающего алгоритма вообще не может быть оценено заранее.

Таким образом, выбрав множества, где число шагов при распознавании находится в указанной зависимости от длины цепочки, мы сразу выделили в классе рекурсивных множеств достаточно узкий подкласс.

---

*) Здесь используется известная формула числа сочетаний:

$$C_P^k = \frac{P!}{k!\,(P-k)!}.$$

**) Под «шагом» здесь подразумевается построение очередного вывода вместе с проверкой, не оканчивается ли он на x.

Грамматики непосредственно составляющих (НС-грамматики)

Для удобства изучения неукорачивающих грамматик целесообразно ввести еще одно ограничение, которое не изменяет класса порождаемых языков, но зато делает правила грамматики более простыми и единообразными по структуре, а потому обеспечивает элементарность каждого шага в выводе. Именно, потребуем, чтобы в каждом правиле $X \to Y$ левая часть (X) имела вид Z_1CZ_2, где C — в точности один символ (удобно согласиться, чтобы этот символ был всегда вспомогательным), а правая часть (Y) — вид Z_1WZ_2, где W — непустая цепочка (непустота цепочки W следует из того, что грамматика является неукорачивающей). Таким образом, на каждом шаге вывода разрешается заменять только один символ. Грамматики, удовлетворяющие сформулированному ограничению, называются грамматиками непосредственно составляющих, или сокращенно — НС-грамматиками. Понятно, что шаг вывода в неукорачивающей грамматике, состоящий в одновременной замене нескольких символов, может быть разбит на несколько шагов, состоящих каждый в замене только одного символа. Отсюда видно, что для любой неукорачивающей грамматики может быть построена эквивалентная ей НС-грамматика *). $T_{1.1.2}$

*) Не приводя полного доказательства этого утверждения, поясним его идею на частном случае правила вида $AB \to BA$, где A и B — вспомогательные символы. Такое правило может быть заменено четырьмя НС-правилами:

$$AB \to 1B$$
$$1B \to 12$$
$$12 \to B2$$
$$B2 \to BA$$

(здесь 1 и 2 — новые вспомогательные символы, которые не встречаются ни в каких старых правилах). Ясно, что последовательное применение этих правил равносильно применению правила $AB \to BA$, причем замена ими этого последнего не может привести к появлению «лишних» выводов, поскольку символы 1 и 2 — новые.

НС-грамматики обладают следующим важным в лингвистическом аспекте свойством.

Будем истолковывать терминальные символы как словоформы (некоторого естественного языка), вспомогательные символы — как синтаксические категории (например, V — глагол, S — существительное, A — прилагательное, $\widetilde{V}$ — группа глагола, $\widetilde{S}$ — группа существительного), начальный символ — как «Предложение», а выводимые терминальные цепочки — как правильные предложения данного языка. Тогда вывод предложения естественно интерпретируется как его синтаксическая структура, представленная в терминах непосредственно составляющих, т. е. способом, давно известным в лингвистике. Поясним сказанное примером. Построим грамматику Γ_1, которая будет порождать некоторые русские фразы, синтаксически однотипные и очень простые. Мы выпишем только схему этой грамматики; ее терминальными символами будут русские словоформы, а вспомогательный словарь содержит вышеназванные синтаксические категории. Символы этих категорий снабжены индексами, соответствующими их морфологическим признакам, например $S_{ж,ед,род}$. Начальный символ обозначается через «ПРЕДЛ» (Предложение).

Схема грамматики Γ_1*)

Следует иметь в виду, что каждая строка в данной схеме представляет собой не одно правило, а сокращенную запись нескольких правил (ср. выше, стр. 39—40). Так, строка II.1 представляет 648 правил:

$$\widetilde{S}_{м,\,ед,\,им,\,3} \to \widetilde{S}_{м,\,ед,\,им,\,3}\,\widetilde{S}_{м,\,ед,\,род,\,1}$$

$$\widetilde{S}_{м,\,ед,\,род,\,3} \to \widetilde{S}_{м,\,ед,\,род,\,3}\,\widetilde{S}_{м,\,ед,\,род,\,1}$$

. . . .

$$\widetilde{S}_{ср,\,мн,\,предл,\,3} \to \widetilde{S}_{ср,\,мн,\,предл,\,3}\,\widetilde{S}_{ср,\,мн,\,род,\,3}$$

*) Смысл используемых обозначений разъясняется ниже, после схемы грамматики.

Такой же способ сокращения применяется и в последующих примерах. Тем не менее для простоты формулировок мы будем называть строки таких сокращенных записей «правилами».

I. Выбор общей структуры предложения.

$$\text{ПРЕДЛ} \to \# \widetilde{S}_{x,\ y,\ \text{им},\ w} \widetilde{V}_{y,\ \text{наст},\ w} \#$$

II. Развертывание именной группы.

1. $\widetilde{S}_{x,\ y,\ z,\ 3} \to \widetilde{S}_{x,\ y,\ z,\ 3} \widetilde{S}_{x',\ y',\ \text{род},\ w}$
2. $\widetilde{S}_{x,\ y,\ z,\ 3} \to A_{x,\ y,\ z} \widetilde{S}_{x,\ y,\ z,\ 3}$
3. $K_1 \widetilde{S}_{x,\ y,\ z,\ w} K_2 \to K_1 S^{\text{мест}}_{x,\ y,\ z,\ w} K_2$

где K_1 — символ, отличный от символа $A_{x,\ y,\ z}$, а K_2 — символ, отличный от символа с индексом z' = род. Символы K_1 и K_2 являются здесь контекстными ограничениями. Содержательный смысл их введения в данное правило заключается в том, что главный член именной группы не должен реализовываться личным местоимением, если ему предшествует определение, выраженное согласованным прилагательным, или если за ним следует именная группа в род. падеже; ср. невозможность **новый я* *) или **он нежности*.

4. $\widetilde{S}_{x,\ y,\ z,\ 3} \to S_{x,\ y,\ z}$

III. Развертывание глагольной группы.

1. $\widetilde{V}_{y,\ \text{наст},\ w} \to V_{y,\ \text{наст},\ w} \widetilde{S}_{x',\ y',\ \text{дат},\ w'} \widetilde{S}_{x'',\ y'',\ \text{твор},\ w''}$
2. $\widetilde{V}_{y,\ \text{наст},\ w} \to V_{y,\ \text{наст},\ w} \widetilde{S}_{x',\ y',\ \text{твор},\ w'} \widetilde{S}_{x'',\ y'',\ \text{дат},\ w''}$
3. $\widetilde{V}_{y,\ \text{наст},\ w} \to V_{y,\ \text{наст},\ w} \widetilde{S}_{x',\ y',\ \text{дат},\ w'}$
4. $\widetilde{V}_{y,\ \text{наст},\ w} \to V_{y,\ \text{наст},\ w} \widetilde{S}_{x',\ y',\ \text{твор},\ w'}$

---

*) В поэтической речи подобные сочетания допускаются: *в четырехлетнюю меня* (М. Цветаева).

IV. Реализация синтаксических категорий словоформами.

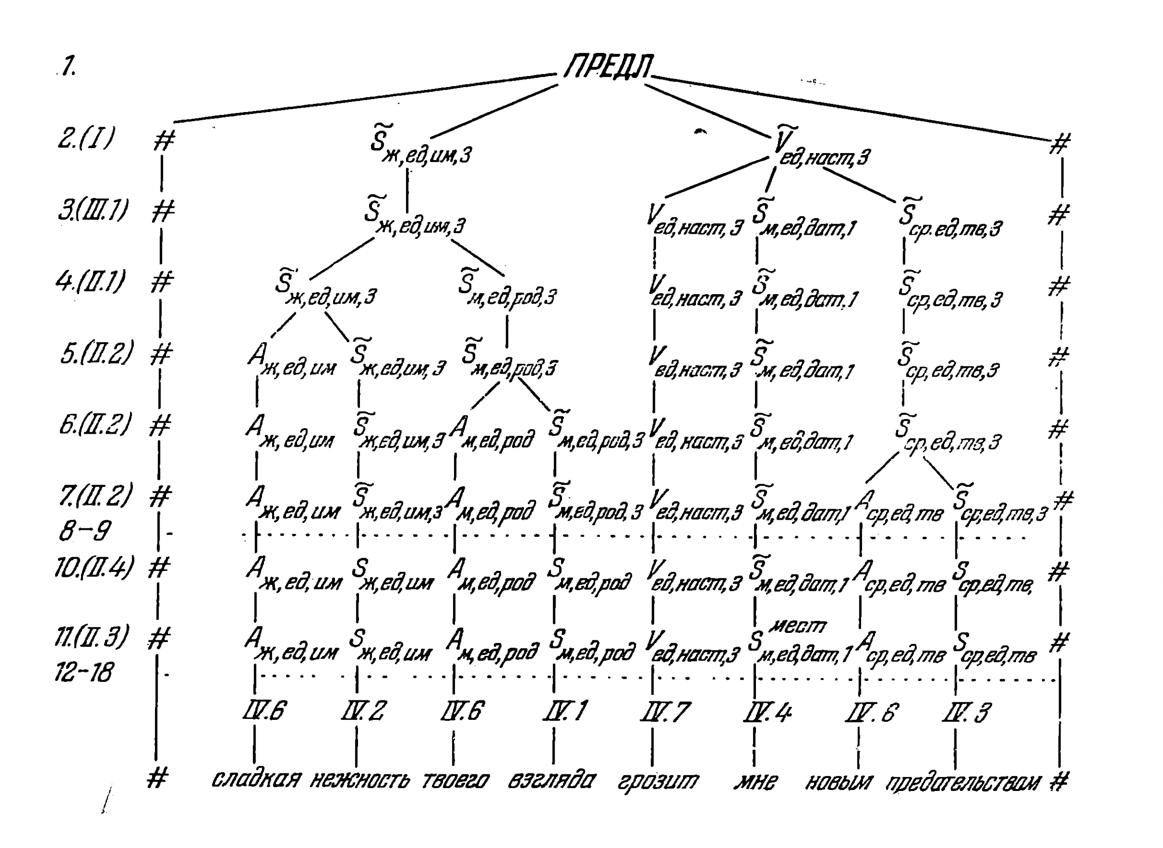
1. $S_{\text{м}, y, z} \rightarrow$ *взгляд*$_{y, z}$, ...
2. $S_{\text{ж}, y, z} \rightarrow$ *нежность*$_{y, z}$, ...
3. $S_{\text{ср}, y, z} \rightarrow$ *предательство*$_{y, z}$, ...
4. $S^{\text{мест}}_{x, \text{ед}, z, 1} \rightarrow$ *я*$_{z}$
5. $S^{\text{мест}}_{x, \text{ед}, z, 2} \rightarrow$ *ты*$_{z}$
6. $A_{x, y, z} \rightarrow$ *сладкий*$_{x, y, z}$, *новый*$_{x, y, z}$, *мой*$_{x, y, z}$, *твой*$_{x, y, z}$, ...
7. $V_{y, \text{наст}, w} \rightarrow$ *грозить*$_{y, \text{наст}, w}$, ...

(В правилах IV не учтено согласование A с одушевленными S в вин. падеже.)

Обозначения: # — символ границы предложения, который считается терминальным (в тексте левая граница реализуется заглавной буквой первого слова, а правая — точкой); x, y, z, w — переменные, соответствующие роду, числу, падежу, лицу (в правилах IV они используются для обозначения словоформ, например, *сладкий*$_{\text{ж,ед,им}}$=*сладкая*).

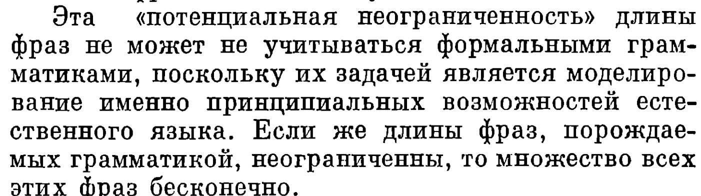
Пример вывода в грамматике Γ_1 см. на стр. 58.

Грамматика Γ_1 способна порождать и другие фразы (которые не обязательно должны быть осмысленными), например: # *я грожу твоей нежности* #, # *сладкое новое мое предательство грозит тобой твоему взгляду предательства твоего взгляда предательства* # и т. п. Более того, грамматика Γ_1 порождает бесконечно много разных фраз (в отличие от грамматики Γ_0). Дело в том, что в ее состав входят так называемые циклические правила —II.1 и II.2; особенность такого правила заключается в том, что результат его применения содержит вхождение его левой части, так что оно всегда может быть применено к своему собственному результату, что и приводит к бесконечному числу фраз: так, наряду с группой *сладкая нежность* можно получить *сладкая сладкая нежность*, далее *сладкая сладкая сладкая нежность* и т. д., т. е. прилагательное *сладкая* может быть повторено сколько угодно раз. В связи с этим


1.
ПРЕДЛ
2.(I)
3.(III.1)
4.(II.1)
5.(II.2)
6.(II.2)
7.(II.2)
8–9
10.(II.4)
11.(II.3)
12–18
IV.6 IV.2 IV.6 IV.1 IV.7 IV.4 IV.6 IV.3
сладкая нежность твоего взгляда грозит мне новым предательством

встает исключительно важный и далеко не простой вопрос о бесконечности числа фраз в естественном языке, относительно которого мы отметим здесь следующее.

Очевидно, что в каждый данный момент число слов любого естественного языка конечно. Кроме того, максимальная длина встречающихся в языке фраз практически ограничена: вряд ли люди употребляют фразы более чем, скажем, в 1000 слов (если это число покажется недостаточным, можно взять любое большее). Отсюда следует, что число фраз в естественном языке должно быть конечно. Тем не менее столь же очевидно, что *указать самую длинную фразу невозможно*: какую бы фразу нам ни предложили, мы всегда в принципе можем удлинить ее, добавив к ней, например, еще один однородный член или предложение с *который*. Это означает, что в естественном языке существуют принципиальные возможности для построения сколь угодно длинных фраз, т. е. потенциально осуществимы фразы любой длины, хотя на практике слишком большие фразы не используются.

Эта «потенциальная неограниченность» длины фраз не может не учитываться формальными грамматиками, поскольку их задачей является моделирование именно принципиальных возможностей естественного языка. Если же длины фраз, порождаемых грамматикой, неограниченны, то множество всех этих фраз бесконечно.

Кроме того, когда мы рассматриваем множество фраз как бесконечное, правила обращения с ними получаются более однотипными, а описание в целом — как это ни парадоксально на первый взгляд — более простым и более глубоким, т. е. позволяющим вскрыть более существенные закономерности. Совершенно аналогичная ситуация имеет место, например, в арифметике: натуральные числа, с которыми людям приходится иметь дело в каких бы то ни было практических задачах, не бывают «слишком большими» — так, навряд ли когда-либо было использовано число, большее 10^{100}. Тем не менее

арифметика исходит из неограниченности натурального ряда (ведь принципиально он и в самом деле ничем не ограничен), и именно это делает ее законы весьма общими и простыми, а потому удобными для применения и в тех задачах, в которых «слишком большие» числа не встречаются.

Что же касается практической ограниченности длины фраз, то это обстоятельство само по себе очень важно и полная теория языка, безусловно, должна его учитывать. Однако природа этого явления относится к иному аспекту языка — не к тому (синтаксическому) аспекту, моделировать который призваны формальные грамматики; поэтому в рамках этих последних оно и не отражается.

Разумеется, в результате истинная картина существенно огрубляется. Однако здесь нет ничего неожиданного; напомним, что во Введении (стр. 10) специально отмечалась неизбежность огрубления при математическом моделировании. Именно так обстоит дело и с формальными грамматиками, причем огрубление происходит здесь не только из-за неучета практической ограниченности длины фраз, но и в гораздо большей степени, например, из-за полного игнорирования семантического аспекта, в частности семантической сочетаемости. Чтобы отразить эти, а также и другие нужные аспекты, необходимо строить наряду с формальными грамматиками модели иного типа, специально приспособленные для описания таких аспектов. (О попытках моделирования семантики см. ниже, стр. 153 и сл.)

Вернемся теперь к нашему примеру вывода на стр. 58. Каждый шаг этого вывода состоит либо в развертывании одного из символов предыдущей цепочки (так, при переходе от цепочки 2 к цепочке 3 символ $\widetilde{V}_{\text{ед,наст,3}}$ развертывается в три символа — $V_{\text{ед,наст,3}}$ $\widetilde{S}_{\text{м,ед,дат,1}}\widetilde{S}_{\text{ср,ед,твор,3}}$), либо в замене его другим (например, при переходе от цепочки 10 к цепочке 11 символ $\widetilde{S}_{\text{м,ед,дат,1}}$ заменяется на $S^{\text{мест}}_{\text{м,ед,дат,1}}$), прочие же символы переписываются *) без изменения.

*) См. Приложение II, стр. 186, Правила подстановки.

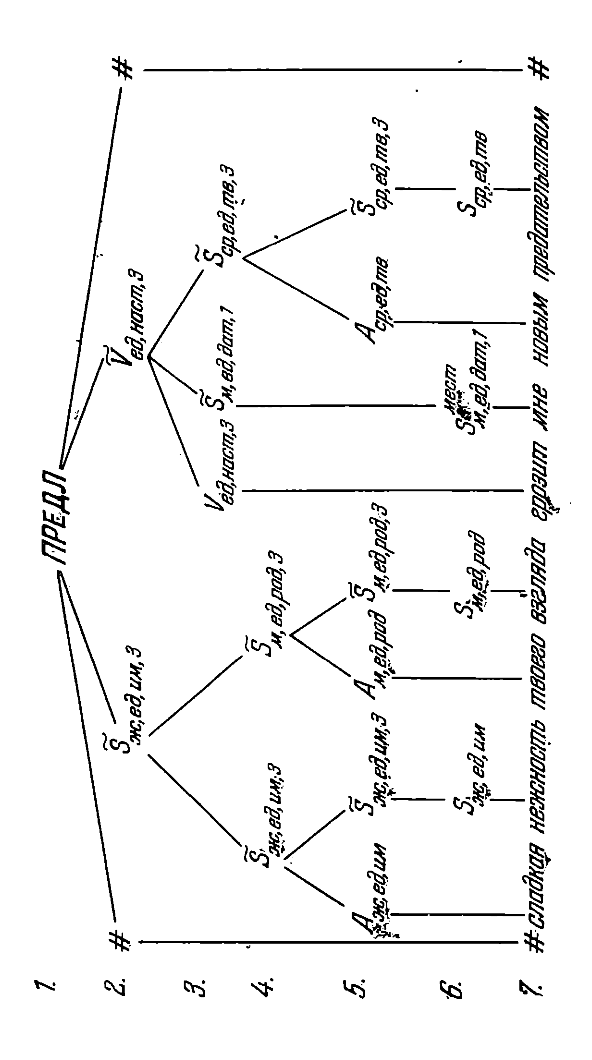
Назовем развертываемые, заменяемые или переписываемые символы «предками», а символы, которые получаются в результате развертывания, замены или переписывания — их «потомками» (потомки потомков также суть потомки). Соединим предков линиями с их непосредственными потомками. Тогда у нас получится не что иное, как хорошо знакомое лингвистам дерево составляющих, или синтаксическая структура фразы в терминах непосредственно составляющих (НС-структура) *). Чтобы показать это более явно, устраним из схемы на стр. 58 все символы-потомки, переписываемые без изменения (например, $\widetilde{S}_{\text{м,ед,дат,1}}$ в цепочках 4—10) и объединим однотипные шаги 4—5, 5—6 и 6—7, изобразив их на одном уровне. Получится дерево, изображенное на стр. 62.

Представление синтаксической структуры в терминах НС широко принято в лингвистике, многократно исследовалось в самых разных аспектах и, безусловно, завоевало право гражданства как в чисто теоретическом плане, так и в работах экспериментального характера (автоматический перевод и т. п.). Поэтому тот факт, что грамматики, определенные на стр. 54, при порождении терминальных цепочек, например, фраз естественного языка, о д н о в р е м е н н о д а ю т и х Н С - с т р у к т у р у (вывод фразы фактически и есть ее НС-структура), делает их особо интересными с лингвистической точки зрения.



Произвольная неукорачивающая грамматика (без требования заменять сразу только один символ) уже не обладает свойством сопоставлять фразам их НС-структуры. Поскольку в такой грамматике каждый раз заменяется, вообще говоря, не один символ, а целая группа символов, в выводе невозможно однозначно указать для каждого символа его предка, и поэтому вывод не может быть превращен в НС-структуру.

*) О составляющих см. также ниже, стр. 164 и сл.


1.
2.
3.
4.
5.
6.
7.
ПРЕДЛ
#
#
сладкая нежность твоего взгляда грозит мне новым предательством

Контекстно-свободные грамматики (КС-грамматики)

Обратим внимание на то, что, хотя в правилах НС-грамматики заменяется только один символ (C, стр. 54), левая часть правила (X) не обязательно состоит только из этого символа: в левой части слева и справа от C могут присутствовать другие символы— контекст, т. е. X может иметь вид Z_1CZ_2. Тогда правило вида $Z_1CZ_2 \rightarrow Z_1WZ_2$ означает разрешение заменять C на W только в контексте $Z_1 ... Z_2$. Сам контекст при этой замене переписывается без изменения. Примером правила с использованием контекста является в грамматике Γ_1 правило II.3. Назовем такие правила контекстно-связанными, а правила, не использующие контекста (т. е. правила вида $X \rightarrow Y$, где X—один символ),— контекстно-свободными. Грамматики, содержащие только контекстно-свободные правила, называются контекстно-свободными, или КС-грамматиками*). Языки, порождаемые КС-грамматиками, естественно называть КС-языками.

КС-грамматики представляют собой важный частный случай НС-грамматик. Их ценность обусловлена следующими двумя обстоятельствами. Во-первых, отказ от контекста, т. е. требование, чтобы в левой части правила был ровно один символ, делает структуру грамматик еще более простой, что облегчает их изучение. Во-вторых, хотя в естественных языках замена одних единиц другими часто допустима лишь в определенных контекстах, целесообразно исследовать возможность описывать языки, отвлекаясь от указанного факта. Это позволяет четко разграничить случаи, где использование контекста действительно необходимо, и случаи, где в принципе можно обойтись и без контекста. В частности, особый интерес представляет исследование ситуаций, где контекст содержательно необходим, но формально

*) НС-грамматики, содержащие контекстно-связанные правила, называются соответственно контекстно-связанными грамматиками.

он может быть учтен с помощью контекстно-свободных правил, т. е. перестает рассматриваться как контекст; этого можно достичь введением в грамматику новых категорий. Так, в грамматике $Г_1$ контекстно-связанное правило II.3 может быть устранено, если преобразовать грамматику следующим образом: 1) во вспомогательный словарь вводятся новые символы $\widetilde{S}'_{x,y,z}$, интерпретируемые как неместоименные именные группы, в отличие от символов $\widetilde{S}_{x,y,z}$, обозначающих произвольные именные группы; 2) правило II.3 заменяется двумя новыми правилами: $\widetilde{S}_{x,y,z,w} \rightarrow S^{\text{мест}}_{x,y,w}$ и $\widetilde{S}_{x,y,z,з} \rightarrow \widetilde{S}'_{x,y,z}$; 3) в правилах II.1, II.2 и II.4 все вхождения символов $\widetilde{S}_{x,y,z,з}$ заменяются символами $\widetilde{S}'_{x,y,z,w}$. Содержательно это означает вот что: при развертывании произвольной именной группы $\widetilde{S}_{x,y,z,w}$ в конструкцию $A + S$ или $S + S_{\text{род}}$ надо следить за тем, чтобы в позиции главы конструкции не оказалось личное местоимение типа *я*, *вы*, *он*, которое не может иметь при себе определений (A или $S_{\text{род}}$: **новый я* или **мы взгляда*). Это может быть обеспечено разными способами. Один из них использован в грамматике $Г_1$: личные местоимения считаются существительными (хотя и особого класса — $S^{\text{мест}}$) и рассматриваются как именные группы ($\widetilde{S}$) наравне с «обычными» существительными; однако переходить от именной группы $\widetilde{S}$ к $S^{\text{мест}}$ разрешается лишь при условии, что это $\widetilde{S}$ раньше не «выделило из себя» A влево или $S_{\text{род}}$ вправо (см. правила II.1 и II.2), т. е. если слева от заменяемого символа нет прилагательного, а справа нет группы существительного в род. падеже. Это условие учитывается правилом II.3. При другом способе (см. выше) местоимения также считаются особым классом существительных, однако наряду с категорией «произвольная именная группа» $\widetilde{S}$ вводится категория собственно именной (неместоименной) группы $\widetilde{S}'$, и символ $\widetilde{S}$ в ходе вывода — до его развертывания — обязательно заменяется либо на символ $S^{\text{мест}}$ (который не может развертываться далее),

либо на символ $\widetilde{S}'$ (который развертывается обычным образом); A и $S_{\text{род}}$ могут появиться только из $\widetilde{S}'$, однако $\widetilde{S}'$ не может превратиться в местоимение. Возможен и третий способ: не считать местоимения существительными и с самого начала вывода использовать для них особый символ M; тогда многие правила грамматики Γ_1 придется продублировать, например, наряду с правилом I ввести правило I′: ПРЕДЛ $\rightarrow M_{\text{ед,им},w}\widetilde{V}_{\text{ед,наст},w}$; наряду с правилом III.3 — правило III.3′: $\widetilde{V}_{y,\text{наст},w} \rightarrow V_{y,\text{наст},w} M_{x',y',\text{дат},w'}$, и т. д. При третьем способе полученная грамматика также будет контекстно-свободной.

Таким образом, разобранный пример показывает, что в естественных языках возможны ситуации, когда явления, которые представляются существенно зависящими от контекста, могут, по-видимому, описываться и как не зависящие от контекста, т. е. в терминах КС-грамматик. При этом, разумеется, описание может усложниться в других отношениях, например, может понадобиться много новых категорий и/или правил. В каждом отдельном случае надо решать, что предпочтительнее, исходя из конкретной задачи описания.

Не следует, однако, думать, что в с я к а я контекстно-связанная НС-грамматика может быть заменена эквивалентной КС-грамматикой. Известно, что существуют НС-языки, не являющиеся КС-языками, например, язык, состоящий в точности из всевозможных цепочек вида $a^n b^n a^n$ (aba, $aabbaa$,...) или из всевозможных цепочек вида $a^n b^n c^n$. **$T_{1.1.3}$**

Почти все имеющиеся примеры НС-языков, не являющихся КС-языками, носят абстрактный характер и не имеют интерпретаций в естественных языках. Ср., впрочем, примеры на стр. 92 и сл.

Итак, до сих пор мы занимались введением все новых и новых ограничений на классы рассматриваемых грамматик. Сначала мы потребовали, чтобы число символов в правой части правил было не меньше, чем в левой, и получили неукорачивающие грамматики. Затем мы потребовали, чтобы замене подвер-

гался только один символ, и получили НС-грамматики. Наконец, мы потребовали, чтобы в левой части правила вообще был только один символ, и получили КС-грамматики. Ясно, что никаких дальнейших естественных ограничений на левые части правил наложить уже нельзя. Поэтому, если мы хотим выделить еще более узкие классы грамматик (о целесообразности такого подхода речь пойдет ниже), придется накладывать ограничения на правые части правил.

Бинарные КС-грамматики

Начнем с числа символов: потребуем, чтобы правая часть любого правила содержала не более двух символов (очевидно, что два — это минимальное число: если в правых частях допускать только по одному символу, порождаемый язык будет состоять из одноэлементных цепочек, что не интересно). В результате получится частный класс КС-грамматик (назовем их бинарными), обладающих той особенностью, что в соответствующих им деревьях составляющих (т. е. в синтаксических структурах фраз, получаемых из их выводов) из каждой вершины исходит не более двух ветвей. Это значит, что любая сложная составляющая всегда состоит ровно из двух непосредственно вложенных в нее составляющих, т. е. что фраза членится всегда на две половины (например, «группа подлежащего» + «группа сказуемого»), каждая из этих половин опять членится пополам и т. д. НС-грамматики (в частности, КС-грамматики) без данного ограничения бинарной структуры не дают. Мы обращаем внимание читателя на этот факт потому, что существует тенденция понимать НС-структуру обязательно как бинарную. Так, грамматика Γ_1 не дает бинарной структуры. Это происходит благодаря правилам III.1 и III.2, которые отражают такое интуитивное понимание строения предложения, при котором группа сказуемого считается состоящей из как бы лежащих на одном уровне личного глагола и групп его дополнений.

Однако для всякой КС-грамматики можно построить эквивалентную ей бинарную КС-грамматику. Например, КС-грамматика, описанная на стр. 64, $\mathbf{T}_{1.1.4.}$ может быть превращена в бинарную путем замены правил III.1 и III.2 (стр. 56) следующими новыми правилами:

$$\text{III. } 1'.\ \widetilde{V}_{y,\ \text{наст},\ w} \to \widetilde{V}^{1}_{y,\ \text{наст},w}\ \widetilde{S}_{x'',\ y'',\ \text{твор},\ w''}$$

$$\text{III. } 2'.\ \widetilde{V}_{y,\ \text{наст},\ w} \to \widetilde{V}^{2}_{y,\ \text{наст},w}\ \widetilde{S}_{x'',\ y'',\ \text{дат},\ w''}$$

$$\text{III. } 1''.\ \widetilde{V}^{1}_{y,\ \text{наст},\ w} \to V_{y,\ \text{наст},w}\ \widetilde{S}_{x',\ y',\ \text{дат},\ w'}$$

$$\text{III. } 2''.\ \widetilde{V}^{2}_{y,\ \text{наст},\ w} \to V_{y,\ \text{наст},w}\ \widetilde{S}_{x',\ y',\ \text{твор},\ w'}$$

Кроме того, необходимо еще заменить правило I, имеющее в правой части четыре символа, правилом I′:

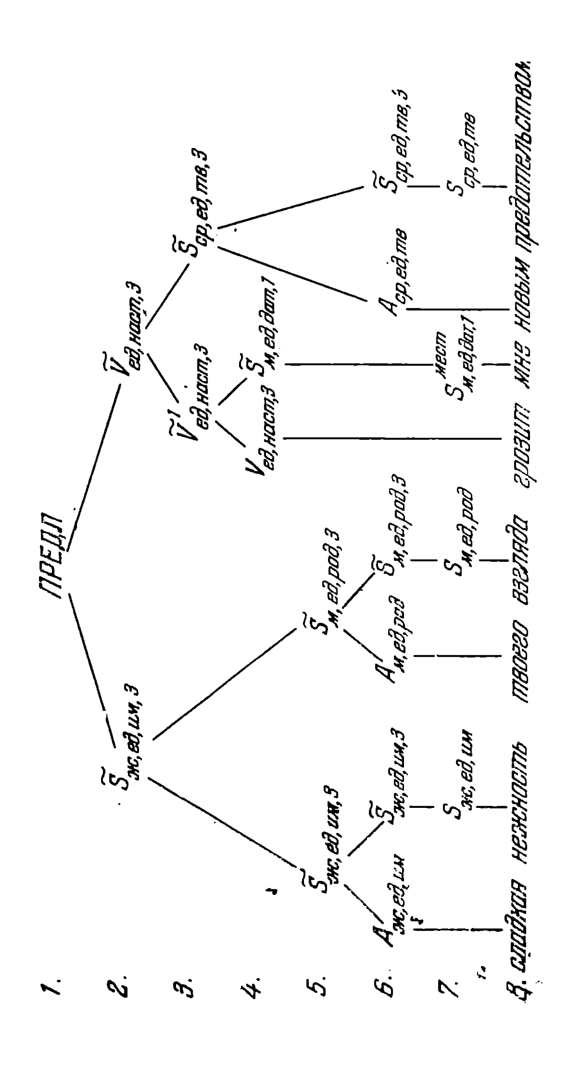
$$\text{ПРЕДЛ} \to \widetilde{S}_{x,\ y,\ \text{им},\ w}\widetilde{V}_{y,\ \text{наст},\ w};$$

тем самым мы устраняем граничные символы (вообще в КС-грамматике граничные символы формально не нужны, тогда как в НС-грамматике, имеющей контекстно-связанные правила, граничные символы могут быть необходимы в качестве контекста: ср. правило II.3 в Γ_1; если же наличие граничных символов считается желательным из содержательных соображений, то их можно сохранить и в бинарной КС-грамматике, однако ради простоты изложения мы этого делать не будем).

Вывод фразы с помощью грамматики, преобразованной указанным образом, будет давать именно бинарную НС-структуру, представленную на стр. 68.

Введенное ограничение (не больше двух символов в правой части правил) можно наложить и на произвольную НС-грамматику, формулируя его, однако, несколько по-другому. Именно, потребуем, чтобы каждое правило имело вид $Z_1CZ_2 \to Z_1WZ_2$, где W состоит из одного или двух символов. Такую НС-грамматику естественно также назвать **бинарной**. Нетрудно показать, что и для всякой НС-грамматики можно построить эквивалентную ей бинарную НС-грамматику.





ПРЕДЛ
1.
2.
3.
4.
5.
6.
7.
8.
сладкая нежность твоего взгляда грозит мне новым предательством

Таким образом, любая фраза, описываемая НС-грамматикой (в частности, КС-грамматикой) и получающая в соответствии с этой грамматикой некоторую НС-структуру, всегда может быть описана бинарной НС-(КС-)грамматикой, т. е. ее НС-структура всегда может быть представлена в терминах строго бинарных составляющих; это можно сделать «равномерно», одинаковым образом для всех фраз языка. Сказанное здесь есть доказанный факт; однако отсюда отнюдь не следует, что бинарное представление фраз естественного языка всегда является удовлетворительным, естественным с точки зрения содержательной лингвистической интерпретации. Теория утверждает лишь, что одни и те же явления могут быть описаны по-разному: бинарно и небинарно (сходная ситуация отмечалась выше — контекстно-связанное и контекстно-свободное описания одних и тех же фраз, стр. 64). Критерии выбора подходящего описания лежат вне теории: этот выбор должен делаться на основе соображений, относящихся к конкретным целям и характеру поставленной задачи.

Автоматные грамматики (А-грамматики)

Теперь мы вернемся на наш путь — путь введения дополнительных ограничений на рассматриваемые грамматики. Поскольку число символов в правой части правил уже сделано минимальным, остается накладывать ограничения на характер заменяющих символов. Потребуем, например, чтобы правая часть каждого правила либо состояла из одного символа, либо имела вид «bB», где b — терминальный (основной) символ, а B — вспомогательный символ (синтаксическая категория). В результате мы получим частный класс КС-грамматик; грамматики этого класса называются **автоматными** (**А-грамматиками**)*)

*) Обычно А-грамматики определяют несколько иначе: требуют, чтобы в правилах, содержащих в правой части только один символ, этот символ был терминальным. Однако классы языков, порождаемых А-грамматиками в смысле того и другого определений, совпадают; поэтому мы позволяем себе использовать тот же термин.

Языки, порождаемые А-грамматиками, также называются **автоматными**, или **A-языками**.

Важнейшей особенностью А-грамматик является специфическая форма вывода. Построим для примера А-грамматику Γ_2, имея в виду порождение предложений типа *Сладкая нежность грозит новым предательством* (упрощенный вариант предложения со стр. 58).

Схема грамматики Γ_2

1. ПРЕДЛ $\to S_{x,\, y,\, \text{им}}$
2. $S_{x,\, y,\, z} \to$ *сладкий*$_{x,\, y,\, z}\, S_{x,\, y,\, z}$
3. $S_{x,\, y,\, z} \to$ *новый*$_{x,\, y,\, z}\, S_{x,\, y,\, z}$
4. $S_{\text{ж},\, y,\, \text{им}} \to$ *нежность*$_{\text{ж},\, y,\, \text{им}}\, V_{y,\, 3}$
5. $S_{\text{ср},\, y,\, \text{им}} \to$ *предательство*$_{\text{ср},\, y,\, \text{им}}\, V_{y,\, 3}$
6. $S_{\text{ж},\, y,\, \text{твор}} \to$ *нежность*$_{\text{ж},\, y,\, \text{твор}}$;
7. $S_{\text{ср},\, y,\, \text{твор}} \to$ *предательство*$_{\text{ср},\, y,\, \text{твор}}$
8. $V_{y,\, 3} \to$ *грозить*$_{y,\, 3}\, S_{x,\, y',\, \text{твор}}$

Указанное предложение будет иметь в данной грамматике следующий вывод:

ПРЕДЛ

(1) $S_{\text{ж},\, \text{ед},\, \text{им}}$

(2) *сладкая* $S_{\text{ж},\, \text{ед},\, \text{им}}$

(4) *сладкая нежность* $V_{\text{ед},\, 3}$

(8) *сладкая нежность грозит* $S_{\text{ср},\, \text{ед},\, \text{твор}}$

(3) *сладкая нежность грозит новым* $S_{\text{ср},\, \text{ед},\, \text{твор}}$

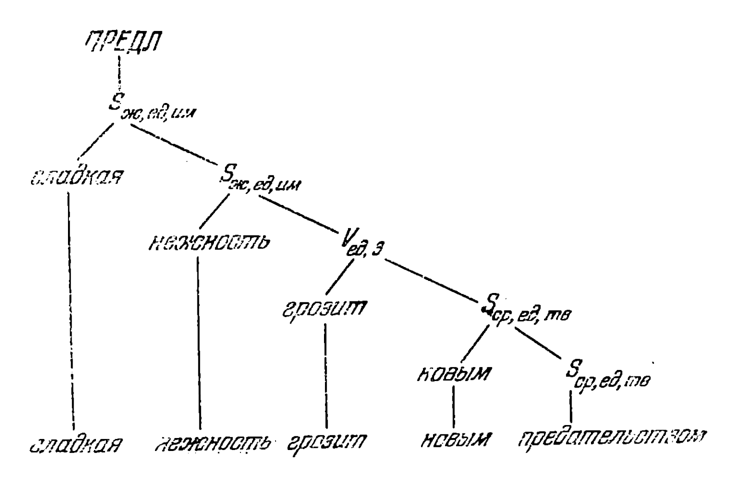
(7) *сладкая нежность грозит новым предательством*

Как мы видим, каждая промежуточная цепочка содержит ровно один вспомогательный символ, стоящий в ней на последнем месте.

Это означает, что предложение порождается слева направо: на каждом шаге выдается конкретная слово-

форма, а за ней — вспомогательный символ, указывающий, какая конструкция должна следовать за этой словоформой; затем (на следующем шаге) выдается словоформа, начинающая или составляющая эту конструкцию, после чего снова следует вспомогательный символ очередной конструкции и т. д. А-грамматика как бы предсказывает, что может следовать за уже выданной словоформой, причем глубина предсказания — один соседний символ; каждый очередной выбор полностью обусловливается только одним предшествующим выбором.

Важно отметить, что из вывода предложения в А-грамматике невозможно получить естественное представление НС-структуры этого предложения (как это делалось для НС- и КС-грамматик). Строго говоря, А-грамматики дают некоторую структуру составляющих, как и вообще все НС-грамматики, однако эти составляющие обычно носят чисто формальный характер и не поддаются естественной интерпретации. Так, приведенный вывод дает разбиение на «составляющие», изображенное ниже.



Вряд ли можно согласиться с разбиением предложения на две составляющие — *сладкая* и все остальное, а также с приписыванием категорий полученным

составляющим. В предложении *Нежность грозит мне предательством* результат был бы еще хуже: составляющей оказалось бы сочетание *мне предательством*! *). Поэтому интерпретация вывода в А-грамматиках как НС-структуры в общем случае не имеет смысла; обычно используется другая интерпретация А-вывода: как последовательности предсказаний и их реализаций.

Наконец, отметим, что класс А-языков у́же класса КС-языков: существуют КС-языки, не порождаемые А-грамматиками. Примером может служить язык, состоящий из всевозможных цепочек вида $a^n b^n$**).

$\mathbf{T}_{1.1.5}$

Заключительные замечания

В предшествующем изложении мы еще далеко не исчерпали всех возможностей ограничивать классы грамматик (в том числе способами, естественными с точки зрения лингвистической интерпретации). Однако других ограничений мы здесь рассматривать не будем; ограничимся тем, что, подводя итог, укажем иерархию выделенных классов грамматик:

ЛВ

произвольные $\supset$ неукорачивающие $\supset$ НС-грамматики $\supset$ КС-грамматики $\supset$ бинарные КС-грамматики $\supset$ автоматные грамматики.

Этим шести классам грамматик соответствует всего четыре класса порождаемых языков:

ЛВ

рекурсивно-перечислимые $\supset$ НС-языки $\supset$ КС-языки $\supset$ А-языки (поскольку неукорачивающие и НС-грамматики порождают одни и те же языки; аналогично, КС-грамматики и бинарные КС-грамматики).

Так как рекурсивно-перечислимые языки — это класс слишком широкий и не представляющий лингвистического интереса, то ниже рассматриваются только три последующих класса, т. е. НС-языки,

---

*) Данное предложение не порождается грамматикой Γ_2, однако читателю будет нетрудно дополнить ее так, чтобы оно порождалось (достаточно добавить два правила).

**) Напомним, что языки $\{a^n b^n a^n\}$ и $\{a^n b^n c^n\}$ не являются КС-языками; ср. стр. 65.

КС-языки и А-языки, а также соответствующие грамматики.

У читателя мог возникнуть вопрос: почему оказывается полезным вводить все новые и новые ограничения на рассматриваемые грамматики, получая все более и более узкие их классы? Это не что иное, как широко принятый в самых разных науках методологический прием: при описании сложного круга явлений сознательно ограничивать набор используемых средств описания, рассматривая и такие средства, которые представляются в общем случае заведомо недостаточными. Исследование может начинаться с самыми минимальными средствами; всякий раз, когда их оказывается недостаточно, постепенно вводятся (и притом возможно более мелкими порциями) новые средства; благодаря этому удается точно определить, какими средствами можно/нельзя обойтись при описании того или иного явления, а тем самым — лучше понять его природу. До появления математических моделей языка подобный подход был в общем чужд лингвистике. Его последовательным внедрением она обязана в первую очередь излагаемой теории грамматик, равно как и математической лингвистике в целом.

§ 4. *Порождающие грамматики и естественные языки*

Теперь мы сделаем некоторые замечания о возможностях описания естественных языков посредством порождающих грамматик. При этом, следуя только что указанному принципу, мы начнем с самого узкого их класса — с автоматных грамматик.

Вопрос о соотношении формальных порождающих грамматик и естественных языков безусловно является с точки зрения лингвиста исключительно важным и интересным; математика, занимающегося теорией формальных грамматик, этот вопрос волнует ничуть не меньше — ведь как раз в естественных языках он должен искать главную содержательную интерпретацию своей теории, а убедительная интерпретация не только повышает «внешнюю» ценность теории, но и служит источником новых идей и методов внутри самой теории.

Однако указанный вопрос весьма сложен и пока что сравнительно мало исследован. Поэтому мы не сможем предложить читателю связного изложения четких результатов, относящихся к описанию конкретных языков с помощью порождающих грамматик того или иного класса. Нам придется ограничиться отдельными соображениями общего характера и немногочисленными известными фактами.

Возможности описания естественных языков с помощью А-грамматик

Для А-грамматики характерно прежде всего, что она, во-первых, порождает фразы строго в одном направлении (при нашем определении — слева направо), развертывая их слово за словом, а во-вторых,

обладает «короткой памятью» — ровно на один шаг. Это означает следующее. Во фразе нередко бывает так, что слова *b* и *c*, далеко отстоящие друг от друга, согласованы в широком смысле этого слова — между ними имеется какое-то определенное соответствие. В КС-грамматике (и тем более в НС-грамматике) это учитывается простым и естественным образом: достаточно, чтобы слова *b* и *c* или их предки появлялись вместе, на одном шаге вывода, как непосредственные потомки одного и того же символа. Именно в этот момент им и приписывается информация о наличии соответствия; после этого между ними может быть вставлено сколько угодно других символов — информация о соответствии все равно сохранится. Так, в грамматике Γ_1 предки подлежащего и личного глагола (символы $\widetilde{S}_{x,y,\text{им},w}$ и $\widetilde{V}_{y,\text{паст},w}$) появляются одновременно как потомки символа ПРЕДЛ при применении правила I; их согласование в числе и лице (y, w) сохраняется до конца вывода, что бы ни было вставлено между ними. Таким образом, информация о согласовании слов *b* и *c* «помнится» при любом расстоянии между ними. В этом смысле можно сказать, что КС-грамматики имеют неограниченную память. Что же касается А-грамматик, то они в том же самом смысле имеют ограниченную память. Дело в том, что А-грамматика способна передавать информацию о соответствии только от непосредственно предшествующего символа к непосредственно следующему: например, в правилах 4 и 5 грамматики Γ_2 (стр. 70) информация о числе (индекс y) передается от подлежащего к непосредственно следующему за ним сказуемому. Поэтому если информацию о соответствии между *b* и *c* приходится передавать через какие-либо промежуточные символы, то в А-грамматике это можно сделать только, приписав указания о наличии соответствия всем промежуточным символам, для которых эти указания по существу не нужны. Так, если мы захотим порождать с помощью А-грамматики русские фразы, где подлежащее отделяется от личного глагола какими-либо словами, например существительным

в родительном падеже (*нежность взглядов грозит предательством*), нам придется ввести специальные правила, где глагол формально будет согласовываться с непосредственно предшествующим ему существительным в родительном падеже. Однако такое согласование будет выполняться не на основе собственных признаков этого существительного, а на основе искусственных «несобственных» признаков, содержательно означающих число (и род) подлежащего — ведь по сути именно с ним и должен согласовываться глагол. Это означает, что вместо каждого символа $S_{x,\,y,\,\text{род}}$ надо будет ввести шесть новых символов: $S_{x,\,y,\,\text{род}\,\|\,\text{м. ед}}$, $S_{x,\,y,\,\text{род}\,\|\,\text{м. мн}}, \ldots$, т. е. «существительное рода x, в числе y, в род. пад., зависящее от подлежащего муж. рода в ед. числе», «существительное рода x, в числе y, в род. пад., зависящее от подлежащего муж. рода во мн. числе» и т. д. Аналогично придется поступить и в случае, когда подлежащее и сказуемое разделяются] наречиями, например, *Нежность взгляда сурово грозит мне*: здесь надо будет ввести специальные категории наречий, а именно: 1) наречия, зависящие от глагола, согласованного с подлежащим муж. рода в ед. числе, 2) наречия, зависящие от глагола, согласованного с подлежащим муж. рода во мн. числе, и т. д.

Более того, если подлежащее и сказуемое разделены одним или несколькими придаточными предложениями, то каждая из категорий, встречающихся в этих предложениях, должна быть расщеплена на шесть категорий, и, стало быть, все правила порождения придаточных предложений должны быть фактически повторены шесть раз. Разумеется, все сказанное сохраняет силу и для тех случаев, когда подлежащее и сказуемое разделены многими словами, например, сколь угодно длинной цепочкой род. падежей или любым числом придаточных предложений. Таким образом, хотя А-грамматика, как и КС-грамматика, способна обеспечивать согласование между сколь угодно далеко отстоящими словами: $\overset{\frown}{a \ldots a'}$ однако это делается громоздким и,

главное, весьма неестественным способом: приходится вводить много дополнительных категорий (классов слов), явно противоречащих языковой интуиции.

Но это еще не все. Если приходится иметь дело не с одной, а со многими согласованными парами, причем каждая следующая «вложена» в предыдущую и число таких пар теоретически не ограничено: $a\ b\ c\ d \cdots d'\ c'\ b'\ a'$, то обеспечить согласование в подобной ситуации А-грамматика принципиально не способна. Этот факт может быть строго доказан (см., например, Гладкий 1966, стр. 103, замечание о языке L_7); излагать самого доказательства мы не будем, а ограничимся содержательными замечаниями, поясняющими, почему это так. Как мы уже видели, для обеспечения согласования одной пары приходится расщепить все категории, разделяющие эту пару; в нашем примере с подлежащим и сказуемым число категорий ушестерялось. Если же между согласуемыми словами имеется еще одна согласуемая пара $a\ b \cdots b'\ a'$, то каждую из уже расщепленных промежуточных категорий надо будет расщепить еще раз, что снова увеличит число категорий (каждая категория, встречающаяся между b и b', должна будет нести указания о согласовании как a с a', так и b с b'). Следовательно, если число вложенных пар потенциально ничем не ограничено, то для обеспечения согласования нужно было бы иметь бесконечно много категорий (вспомогательных символов) — тогда как число символов в любой грамматике конечно. Таким образом, указанная ситуация А-грамматикой описываться не может.

Между тем эта ситуация достаточно типична для естественных языков. Так, она имеет место в сложных предложениях вроде: *Человек*, *которому* (a) *гости* (b), *собравшиеся* (c), *когда* *хозяин* (d), *который* ..., *встал* (d'),

танцевать$^{c'}$ *в соседней комнате, испортили*$^{b'}$ *костюм, ушел*$^{a'}$ *домой* (на место многоточия можно вставить любые предложения, в том числе содержащие много придаточных, также с последовательным вложением). Подобные предложения регулярно встречаются в самых разных языках. Примеры потенциально неограниченного вложения согласуемых пар можно указать и для простых предложений. Таковы, например, русские конструкции с последовательным вложением препозитивных причастных оборотов (например, ... *для расставшейся*^{a} *с разыскивающим*^{b} *влюбленного*^{c} *в эту чарующую*^{d} *всех девушку*$^{d'}$ *человека*$^{c'}$ *писателем*$^{b'}$ *жены*$^{a'}$...) или пар однородных существительных, согласованных в падеже (*от леммы требуется эквивалентность*^{a} *теореме*^{b} *о компактности*^{c} *отрезков*^{d} *и квадратов*$^{d'}$, *или о* Δ-*компактности*$^{c'}$, *но не утверждению*$^{b'}$ *4377, а также простота*$^{a'}$) *), или конструкции типа S' *de* S'' *de* S''' ... $A'''A''A'$ во французском и других романских языках (... *chez la maîtresse*^{a} *d'un membre*^{b} *de Société*^{c} *Linguistique*$^{c'}$ *enrhumé*$^{b'}$ *envoyée*$^{a'}$ *à*... Известны также еще два примера неприменимости А-грамматик (в этих случаях неприменимы и КС-грамматики, см. ниже, стр. 92 и сл.), относящиеся, впрочем, к весьма ограниченным по распространенности явлениям. Что же касается только что приведенных примеров, то, хотя круг языков, в которых допустимы подобные конструкции, вероятно, ограничен, они все же представляются вполне типичными, так что полностью отвлечься от них, по-видимому, нельзя. Поэтому приходится признать, что построить п о л н о е описание естественного

---

*) На конструкции обоих указанных типов обратил внимание авторов Ф. А. Дрейзин.

языка на основе только А-грамматик невозможно. Строго говоря, это означает следующее: либо построенная нами А-грамматика не будет порождать некоторых правильных фраз (в частности, фраз вида $abcd \ldots d'c'b'a'$, см. выше), либо, если мы сделаем ее способной порождать любую правильную фразу (а это всегда возможно!), то она обязательно начнет порождать и некоторые неправильные фразы (например, наряду с фразой $abcd \ldots d'c'b'a'$ она будет порождать и фразу $abcd \ldots c'a'd'b'$ — с нарушенным согласованием).



В дальнейшем, когда речь идет о невозможности описать язык с помощью той или иной грамматики, везде имеется в виду именно это — либо грамматика не порождает некоторых правильных фраз достаточно обычного и распространенного типа (т. е. является неполной), либо обязательно порождает помимо всех правильных фраз и некоторые неправильные (т. е. является неадекватной).

Однако то, что А-грамматики недостаточны для описания естественного языка во всем объеме, еще не исключает возможности описывать с помощью А-грамматик те или иные ф р а г м е н т ы естественного языка. При этом можно, как кажется, предполагать, что в естественном языке «А-фрагмент», как правило, покрывает главную часть. В самом деле, конструкции с неограниченным вложением согласуемых пар немногочисленны, а конструкции типа указанных на стр. 92 к тому же еще и периферийны. Тем самым А-грамматики в принципе способны описывать достаточно существенную часть множества предложений (простых и сложных) естественного языка. Кроме того, А-грамматики могут описывать и другие языковые объекты: например, словосочетания (см. ниже, на стр. 81—83, описание элементарных именных групп), словоформы, слоги. Разумеется, из сказанного вовсе не следует, что во всех тех случаях, когда А-грамматики применимы, они описывают свой объект естественным образом, т. е. что они всегда удобны. Более того, из изложения на стр. 75 — 76 видно, что это

не так. Однако любая грамматика, специально приспособленная для естественного описания какого-либо А-фрагмента языка (скажем, для описания простых предложений, не содержащих конструкций вроде указанных выше), будет эквивалентна некоторой А-грамматике. А так как грамматики, эквивалентные А-грамматикам (хотя и не являющиеся таковыми), обычно в каком-то отношении — либо по характеру правил, либо по характеру выводов — характеризуются приблизительно такой же степенью простоты, что и сами А-грамматики, то тем самым мы получаем как бы эталон простоты. Создавая грамматику, описывающую простые предложения, мы должны стремиться к тому, чтобы она хотя бы в одном из указанных отношений не была намного сложнее, чем А-грамматика. (Напомним, что «простоту» следует понимать в логическом, а не в бытовом смысле! — сноска на стр. 25.) Ср. ниже, на стр. 99 и 103, примеры грамматик с ограниченной памятью и «левоправых» грамматик.

Что же касается вопроса о том, где именно А-грамматики оказываются не только применимыми, но и естественными, то в целом он до сих пор остается малоисследованным, и его выяснение представляет большой интерес. В предварительном порядке можно, по-видимому, считать, что А-грамматики в достаточной степени удобны при описании элементарных именных групп *) типа *или не только из всех этих трех наших первых огромных трехарочных металлических железнодорожных мостов* (здесь приведен пример максимальной «схемы» элементарной именной группы — ЭИГ; в действительности такие группы обычно выступают в более простом виде — те или иные места могут быть не заполнены).

Приведем пример А-грамматики Γ_3, порождающей любые ЭИГ указанного типа (с неодушевл. S).

---

*) Под элементарной именной группой здесь понимается существительное со всеми его препозитивными согласованными определениями, а также предлогом, (ограничительными) частицами и вводящим всю группу сочинительным союзом,

Схема грамматики Г$_3$

I.

$$\text{ЭИГ} \rightarrow \left\{\begin{matrix} \textit{или} \\ \textit{и} \\ \textit{либо} \\ \cdots \\ \Lambda \end{matrix}\right\} \text{ЭИГ}^1$$

Как и выше (стр. 41), Λ — пустая цепочка; содержательно наличие Λ означает, что соответствующее место в ЭИГ может остаться незаполненным.

II.

$$\text{ЭИГ}^1 \rightarrow \left\{\begin{matrix} \textit{ни} \\ \textit{не} \\ \textit{не только} \\ \textit{отнюдь не} \\ \textit{лишь} \\ \textit{только} \\ \textit{хотя бы} \\ \cdots\cdots \\ \Lambda \end{matrix}\right\} \text{ЭИГ}^2$$

III. 1.

$$\text{ЭИГ}^2 \rightarrow \left\{\begin{matrix} \textit{из} \\ \textit{с} \\ \textit{без} \end{matrix}\right\} \text{ЭИГ}^3_{x,\ y,\text{род}}$$

2.

$$\text{ЭИГ}^2 \rightarrow \left\{\begin{matrix} \textit{к} \\ \textit{по} \end{matrix}\right\} \text{ЭИГ}^3_{x,\ y,\ \text{дат}}$$

. . . .

6. $\text{ЭИГ}^2 \rightarrow \text{ЭИГ}^3_{x,\ y,\ z}$

IV.

$$\text{ЭИГ}^3_{x,\ y,\ z} \rightarrow \left\{\begin{matrix} \textit{весь}_{x,\ y,\ z} \\ \textit{каждый}_{x,\ y,\ z} \\ \textit{какой-либо}_{x,\ y,\ z} \\ \cdots\cdots \\ \Lambda \end{matrix}\right\} \text{ЭИГ}^4_{x,\ y,\ z}$$

V.

$$\text{ЭИГ}^4_{x,\ y,\ z} \rightarrow \left\{\begin{matrix} \textit{этот}_{x,\ y,\ z} \\ \textit{тот}_{x,\ y,\ z} \\ \textit{этот же самый}_{x,\ y,\ z} \\ \cdots\cdots \\ \Lambda \end{matrix}\right\} \text{ЭИГ}^5_{x,\ y,\ z}$$

VI.

1. $$\text{ЭИГ}^5_{x,\ y,\ z} \rightarrow \left\{\begin{matrix} \textit{один}_{x,\ y,\ z} \\ \\ \Lambda \end{matrix}\right\} \text{ЭИГ}^6_{x,\ y,\ z}$$

2.

$$\text{ЭИГ}^5_{x,\ y,\ z} \to \left\{\begin{matrix} \textit{два}_{x,\ z} \\ \textit{три}_z \\ \textit{четыре}_z \end{matrix}\right\} \text{ЭИГ}^6_{x,\ \text{ед},\ \text{род}}$$

Здесь $z =$ им, вин

3.

$$\text{ЭИГ}^5_{x,\ y,\ z} \to \left\{\begin{matrix} \textit{пять}_z \\ \textit{шесть}_z \\ \dots\dots \\ \textit{тысяча девятьсот шестьдесят шесть}_z \end{matrix}\right\} \text{ЭИГ}^6_{x,\ \text{мн},\ \text{род}}$$

Здесь $z =$ им, вин

4.

$$\text{ЭИГ}^5_{x,\ y,\ z} \to \left\{\begin{matrix} \textit{два}_{x,\ z} \\ \dots\dots \\ \textit{пять}_z \\ \textit{шесть}_z \\ \dots\dots \\ \textit{тысяча девятьсот шестьдесят шесть}_z \end{matrix}\right\} \text{ЭИГ}^6_{x,\text{мн},\ z}$$

Здесь $z \neq$ им, вин

VII.

$$\text{ЭИГ}^6_{x,\ y,\ z} \to \left\{\begin{matrix} \textit{мой}_{x,\ y,\ z} \\ \textit{твой}_{x,\ y,\ z} \\ \dots\dots \\ \textit{их} \\ \Lambda \end{matrix}\right\} \text{ЭИГ}^7_{x,\ y,\ z}$$

VIII.

$$\text{ЭИГ}^7_{x,\ y,\ z} \to \left\{\begin{matrix} \textit{первый}_{x,\ y,\ z} \\ \textit{второй}_{x,\ y,\ z} \\ \dots\dots \\ \textit{тысяча девятьсот шестьдесят шестой}_{x,y,z} \\ \dots\dots \\ \Lambda \end{matrix}\right\} \begin{matrix} \text{ЭИГ}^{8i}_{x,\ y,\ z} \\ (1 \leqslant i \leqslant p) \end{matrix}$$

IX. $\text{ЭИГ}^{8i}_{x,\ y,\ z} \to a^i_{x,\ y,\ z}\ \text{ЭИГ}^{8j}_{x,\ y,\ z} \qquad (1 \leqslant j \leqslant i \leqslant p)$

Обозначение $a^i{}_{x,\ y,\ z}$ разъясняется в примечании после грамматики.

X. $\text{ЭИГ}^{8i}_{x,\ y,\ z} \to S_{x,\ y,\ z}$

XI. 1.

$$S_{\text{м}, y, z} \to \begin{Bmatrix} мост_{y, z} \\ нос_{y, z} \\ саквояж_{y, z} \\ \ldots\ldots \end{Bmatrix}$$

2.

$$S_{\text{ж}, y, z} \to \begin{Bmatrix} река_{y, z} \\ губа_{y, z} \\ сумка_{y, z} \\ \ldots \end{Bmatrix}$$

3.

$$S_{\text{ср}, y, z} \to \begin{Bmatrix} озеро_{y, z} \\ ухо_{y, z} \\ ведро_{y, z} \\ \ldots\ldots \end{Bmatrix}$$

Примечание к правилу IX. Символ a^i использован здесь для обозначения конкретных прилагательных i-го класса, причем к одному классу относятся прилагательные, занимающие одну и ту же позицию по отношению к определяемому существительному, а неравенство $i > j$ означает, что прилагательное i-го класса должно стоять дальше от существительного, чем прилагательное j-го класса. Например, прилагательное *французский* (*английский*, *советский*,...) имеет индекс класса меньший, чем у прилагательного *интересный* (*новый*, *ценный*, ...), поскольку выражение *интересные французские журналы* обычнее, чем *французские интересные журналы*. Число таких классов прилагательных обозначено через p.

Таким образом, к грамматике $Г_3$ должен быть приложен список прилагательных, снабженных индексами класса в определенном здесь смысле. Для нашего примера мы возьмем небольшой словарь, содержащий прилагательные пяти классов (см. стр. 84).

Данное разбиение выполнено исключительно в иллюстративных целях и отражает действительную картину весьма огрубленно: законы взаимного размещения прилагательных в действительности не укладываются в рамки линейного упорядочения; кроме того, порядок прилагательных зависит на самом деле от логического акцента («актуального членения» словосочетания) — так, *интересная*

1-й класс	2-й класс	3-й класс	4-й класс	5-й класс
политический	*стальной*	*немецкий*	*белый*	*хороший*
музыкальный	*бумажный*	*сомалийский*	*синий*	*негодный*
литературный	*деревянный*	*чешский*	*желтый*	*замечательный*
математический	*костяной*	*грузинский*	*розовый*	*интересный*
химический	*виноградный*	*норвежский*	*черный*	*отличный*
. . . .			. . .	. . . .

математическая книга ≈ ‘имеется математическая книга, она интересна’, тогда как *математическая интересная книга* ≈ ‘имеется интересная книга, она по математике’. Аналогичным образом, в лингвистической работе, посвященной именным конструкциям некоторого языка, речь может идти об *определительных именных конструкциях*, *субъектных именных конструкциях* и т. п., а в работе об определительных конструкциях мы встретимся скорее с *именными определительными конструкциями*, *глагольными определительными конструкциями* и т. п. При описании таких случаев мы имеем в виду наиболее нейтральный, «обычный» порядок.

Приведем пример вывода в грамматике Г$_3$:

ЭИГ

(I) *и* ЭИГ^{1}

(II) *и хотя бы* ЭИГ^{2}

(III.4) *и хотя бы с* ЭИГ$^{3}_{\text{ж, мн, твор}}$

(IV) *и хотя бы с* ЭИГ$^{4}_{\text{ж, мн, твор}}$

(V) *и хотя бы с* ЭИГ$^{5}_{\text{ж, мн, твор}}$

(VI.4) *и хотя бы с девятью* ЭИГ$^{6}_{\text{ж, мн, твор}}$

(VII) *и хотя бы с девятью нашими* ЭИГ$^{7}_{\text{ж, мн, твор}}$

(VIII) *и хотя бы с девятью нашими* ЭИГ$^{8}{}_{5\,\text{ж, мн, твор}}$

(IX) *и хотя бы с девятью нашими отличными* ЭИГ$^{3_2}_{\text{ж, мн, твор}}$

(IX) *и хотя бы с девятью нашими отличными бумажными* $ЭИГ^{3_2}_{ж, мн, твор}$

(X) *и хотя бы с девятью нашими отличными бумажными* $S_{ж, мн, твор}$

(XI.2) *и хотя бы с девятью нашими отличными бумажными сумками.*



(Еще раз позволим себе заметить, что для читателя будет весьма полезно поупражняться в самостоятельном построении выводов. Это относится, конечно, также ко всем предыдущим и последующим примерам грамматик.)

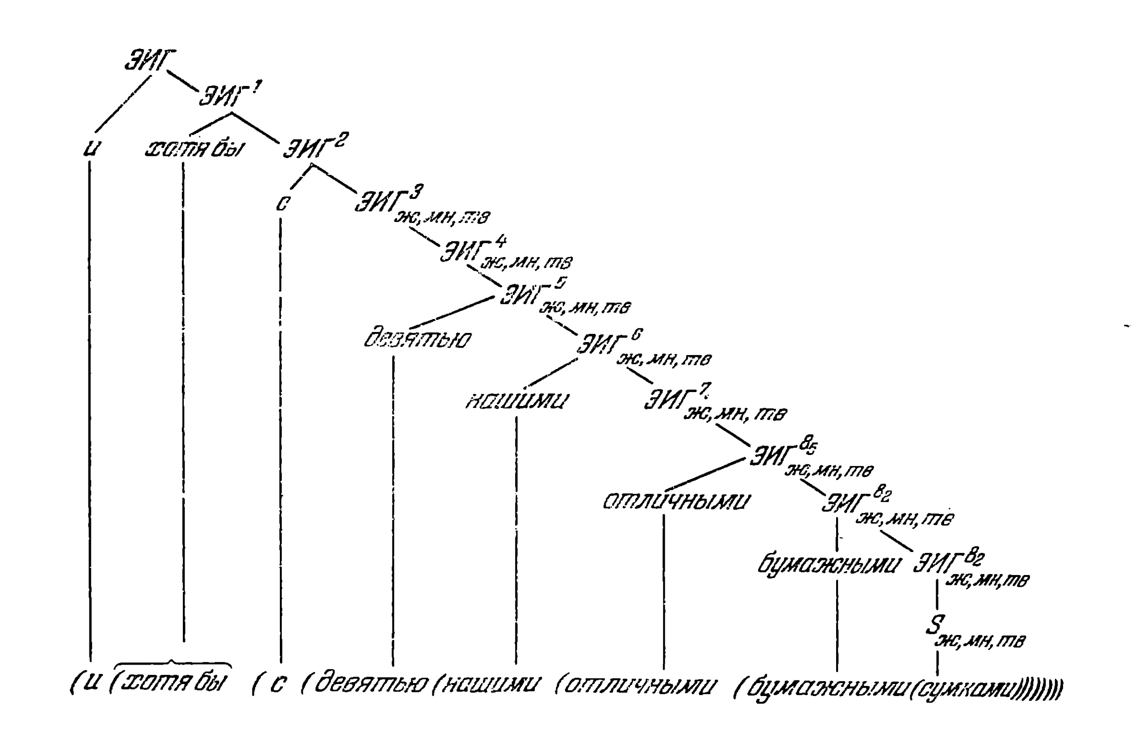
Обратим внимание на то, что составляющие, которые получаются из приведенного вывода способом, описанным на стр. 61, оказываются в данном случае вполне естественными (в отличие от грамматики $Г_2$), см. рис. на стр. 86. Это объясняется особенностями синтаксического строения элементарных именных групп в русском языке — ЭИГ строятся по схеме

$$\ldots (h\,(g\,(f\,(e\,(d\,(c\,(b\,(a)))))))) \ldots,$$

т. е. все элементы, распространяющие главный элемент (a), находятся влево от него, и каждый из них относится (определяет или подчиняет) сразу ко всему, что стоит вслед за ним; при этом любой элемент, кроме главного (т. е. последнего), может отсутствовать. Встречающиеся в выводе вспомогательные символы — ЭИГ¹, ЭИГ², ... — естественно интерпретируются как обозначения для «неполных» ЭИГ, т. е. как типы составляющих: ЭИГ¹ — именная группа без союза, ЭИГ² — именная группа без союза и без ограничительной (отрицательной) частицы, ЭИГ³ — именная группа без союза, без ограничительной (отрицательной) частицы и без предлога, и т. п.

Вообще, А-грамматики в некотором отношении удобны для описания именно тех языковых объектов, которые имеют указанную схему строения, т. е. «наращиваются» только с одной стороны. Такими объектами, например, являются агглютинативно построенные словоформы — вроде таких, как русские

((((((*послед*)*ова*)*тель*)*н*)*ост*)*н*)*ый*)


ЭИГ
ЭИГ¹
и
хотя бы
ЭИГ²
с
ЭИГ³ ж, мн, тв
ЭИГ⁴ ж, мн, тв
ЭИГ⁵ ж, мн, тв
девятью
ЭИГ⁶ ж, мн, тв
нашими
ЭИГ⁷ ж, мн, тв
ЭИГ^{8₅} ж, мн, тв
отличными
ЭИГ^{8₂} ж, мн, тв
бумажными
ЭИГ^{8₂} ж, мн, тв
S ж, мн, тв
(и (хотя бы (с (девятью (нашими (отличными (бумажными (сумками))))))))

или *((((((предсказ)у)ем)ост)н)ост)ь)*, ср. также *((((цел)у)ющ)его)ся)* и т. п. (в языках типа турецкого или венгерского подобные формы носят несравненно более регулярный характер).

Разумеется, для описания объектов, растущих вправо, более естественны А-грамматики не совсем такие, как определенные на стр. 69, а именно: в этом определении следует заменить правила вида $A \to bB$ правилами вида $A \to Bb$, т. е. перейти от левосторонних А-грамматик к правосторонним.

В заключение разъясним, что понимается под удобством применения А-грамматик для описания языковых объектов вида $(((a)b)c)d)\ldots$ — так сказать, «слоистых» объектов. А-грамматики хороши здесь именно потому, что эта слоистость в явной форме вскрывается А-выводом соответствующего объекта (это и есть то самое «некоторое отношение», о котором говорилось на стр. 85). Однако в других аспектах А-грамматики могут и не быть удобными при описании даже таких объектов, например, для отражения всех морфонологических процессов, сопровождающих порождение словоформ. В таких случаях, по-видимому, окажется целесообразным расчленять соответствующие явления на разные уровни и описывать их несколькими грамматиками, одна из которых будет автоматной (например, А-грамматике может быть поручено порождение словоформ на уровне морфем и, может быть, морф, тогда как дальнейшая реализация полученной цепочки выполняется грамматиками других типов).

Теперь мы перейдем к рассмотрению контекстно-свободных (КС)-грамматик.

КС-грамматики и естественные языки

Наиболее характерной особенностью КС-грамматик является тот факт, что на каждом шаге вывода «обрабатывается» р о в н о о д и н символ, т. е. никаким образом не может быть учтено наличие/отсутствие или свойства различных соседних символов.

Это может создать впечатление, что КС-грамматики мало пригодны для описания естественных языков: ведь в обычных грамматиках утверждения о выборе тех или иных форм, о варьировании или развертывании тех или иных элементов высказывания, как правило, формулируются именно с учетом контекстных условий. Так, при описании флективных форм указывается, какая флексия должна быть выбрана в зависимости от типа основы (который, таким образом, выступает как контекст); при описании употребления русских падежей указывается, что винительный падеж прямого дополнения заменяется родительным при наличии отрицания (и некоторых других, более сложных условий); творительный субъекта возможен при отглагольном существительном только в том случае, если при этом существительном есть дополнение в родительном падеже (*рассмотрение вопроса советом*, но не **рассмотрение советом*) и т. д. Однако даже А-грамматика (весьма частный случай КС-грамматики!), как мы видели, практически способна порождать подавляющее большинство простых и сложных предложений естественного языка. Тем более это должно быть верно для произвольных КС-грамматик. В самом деле, оказывается, что почти во всех случаях, где использование контекста на первый взгляд представляется неизбежным, фактически без него в принципе можно обойтись. В самых общих чертах это делается так: пусть имеется класс элементов X, причем в соседстве с элементами некоторого класса Y элементы X ведут себя иначе, чем в соседстве с элементами класса Z, так что, например, имеют место правила:

1. $YX \to YAB$
2. $ZX \to ZCD$ (эти правила используют контекст).

Можно, однако, ввести два новых символа X_1 и X_2 и обозначить через X_1 элемент X в позиции после Y, а через X_2 — элемент X в позиции после Z. Тогда обычно удается перейти к правилам, не использующим контекста:

1′. $X_1 \to AB$,
2′. $X_2 \to CD$,

Другими словами, вводятся более дробные категории элементов, учитывающие их позиции в контексте. Покажем, каким образом может быть выполнен переход к контекстно-свободным формулировкам в приведенных выше примерах обращения к контексту. Слева помещается нужный фрагмент соответствующей контекстно-связанной грамматики, справа — эквивалентный фрагмент, состоящий из контекстно-свободных правил.

а) Выбор флексии падежа в зависимости от типа основы:

$\text{Сф}_{\text{мн, род}} \to \text{О}^i\ \text{Фл}_{\text{мн, род}}$	$\text{Сф}_{\text{мн, род}} \to \text{О}^i\ \text{Фл}^i_{\text{мн, род}}$
$\text{О}^1\,\text{Фл}_{\text{мн, род}} \to \text{О}^1$ *ов* (*штан-ов*)	$\text{Фл}^1_{\text{мн. род}} \to$ *ов*
$\text{О}^2\,\text{Фл}_{\text{мн, род}} \to \text{О}^2$ (*брюк-*)	$\text{Фл}^2_{\text{мн, род}} \to \Lambda$
$\text{О}^3\,\text{Фл}_{\text{мн, род}} \to \text{О}^3$ *ей* (*плащ-ей*)	$\text{Фл}^3_{\text{мн, род}} \to$ *ей*

. . . .

где Сф — словоформа, О^i — основа типа i ($i = 1, 2, 3, \ldots$), $\text{Фл}_{\text{мн,род}}$ — флексия род. падежа мн. числа.

б) Выбор падежа прямого дополнения в зависимости от наличия отрицания:

$\widetilde{V} \to V^{tr} Obj_{\text{dir}}$	$\widetilde{V} \to V^{tr} Obj^1_{\text{dir}}$
$\widetilde{V} \to Neg V^{tr} Obj_{\text{dir}}$	$\widetilde{V} \to Neg V^{tr} Obj^2_{\text{dir}}$
$X V^{tr} Obj_{\text{dir}} \to X V^{tr} \widetilde{S}_{\text{вин}}$	$Obj^1_{\text{dir}} \to \widetilde{S}_{\text{вин}}$
$[X \neq Neg]$	$Obj^2_{\text{dir}} \to \widetilde{S}_{\text{род}}$

(*мальчик читает журнал*)

$$Neg V^{tr} Obj_{\text{dir}} \to Neg V\ \widetilde{S}_{\text{род}}$$

(*мальчик не читает журнала*)

где $\widetilde{V}$ — группа глагола, V^{tr} — переходный глагол, Obj_{dir} — прямое дополнение, $\widetilde{S}$ — группа существительного, Neg — отрицание.

в) Возможность творительного субъекта при отглагольном существительном в зависимости от

наличия объекта (*рассмотрение вопроса советом*):

$$\widetilde{S} \to \widetilde{S}'Obj\ Subj \qquad \widetilde{S} \to \widetilde{S}'Obj\ Subj^1$$
$$\widetilde{S} \to \widetilde{S}'Subj \qquad \widetilde{S} \to \widetilde{S}'Subj^2$$
$$ObjSubj \to Obj\ \widetilde{S}_{\text{твор}} \qquad Subj^1 \to \widetilde{S}'_{\text{твор}}$$

(Obj — объект, $Subj$ — субъект).

Во всех этих трех содержательно весьма различных примерах применен в точности один и тот же формальный прием: информация о контексте «загоняется» в новые категории. Таким образом, чем меньше мы хотим использовать контекст, тем больше категорий приходится ввести, и обратно. Привлекательность перехода к контекстно-свободным правилам состоит в том, что оценить степень сложности разнообразных и содержательно пестрых обращений к контексту очень трудно, тогда как в КС-правилах «мерой» степени сложности становится просто число используемых категорий. Что касается лингвистической осмысленности такого перехода, то в ряде случаев введение новых категорий с отказом от использования контекста оказывается оправданным, а в других случаях может представляться искусственным приемом. Однако же и здесь переход к контекстно-свободным правилам может быть полезен — уже хотя бы тем, что заставляет четко поставить вопрос о содержательной целесообразности использования контекста в конкретных ситуациях.



Итак, как уже было отмечено, п о ч т и во всех случаях контекст может быть элиминирован. Тем не менее есть случаи, когда это не так. В частности, от контекста нельзя отказаться, т. е. невозможно обойтись одним символом в левой части правила, если правило должно обеспечивать перестановку символов: ведь перестановка по своему существу является многоместной операцией. Стало быть, КС-грамматика не может породить язык, содержащий цепочки, которые не могут быть построены без применения перестановок. Рассмотрим, например, язык, содержащий всевозможные цепочки вида $a_1a_2a_3qa'_1a'_2a'_3$,

$a_2a_1a_2a_3qa_2'a_1'a_2'a_3'$, $a_1a_3a_2a_1qa_1'a_3'a_2'a_1'$ и т. п. (в общем виде такие цепочки можно записывать как xqx') и не содержащий никаких других цепочек. Содержательно a_1 и a_1', a_2 и a_2' и т. д. могут пониматься как пары элементов, определенным образом «согласованных» между собой *). Этот язык легко может быть порожден грамматикой, содержащей правила перестановки, например следующей **):

$$\left.\begin{array}{ll} 1.\ I & \to IA_ia_i', \\ 2.\ a_i'A_j & \to A_ja_i', \\ 3.\ IA_i & \to a_iI, \\ 4.\ I & \to q, \end{array}\right\} \quad i, j = 1,\ 2,\ 3$$

(a_i, a_i', q — основные символы; I, A_i — вспомогательные символы; I — начальный символ).

Покажем для примера, как можно вывести в этой грамматике цепочку $a_2a_1a_1a_3qa_2'a_1'a_1'a_3'$:

$$\begin{array}{rl} & I \\ (1) & IA_3a_3' \\ (1) & IA_1a_1'A_3a_3' \\ (1) & IA_1a_1'A_1a_1'A_3a_3' \\ (1) & IA_2a_2'A_1a_1'A_1a_1'A_3a_3' \\ (2) & IA_2A_1a_2'a_1'A_1a_1'A_3a_3' \\ \ldots & \ldots\ldots\ldots\ldots \\ (2;\ 5\ \text{раз}) & IA_2A_1A_1A_3a_2'a_1'a_1'a_3' \\ (3) & a_2IA_1A_1A_3a_2'a_1'a_1'a_3' \\ \ldots & \ldots\ldots\ldots\ldots \\ (3;\ 3\ \text{раза}) & a_2a_1a_1a_3Ia_2'a_1'a_1'a_3' \\ (4) & a_2a_1a_1a_3qa_2'a_1'a_1'a_3' \end{array}$$

---

*) Строго говоря, здесь речь идет не о самих символах a_1, a_1' и т. д., а об их соответствующих вхождениях в цепочки.

**) В соответствии с теоремой $T_{1.1.2}$(стр. 179) для этой грамматики существует эквивалентная НС-грамматика (которую читатель без труда построит сам, пользуясь методом моделирования перестановок НС-правилами, указанным в сноске к стр. 54).

В то же время известно, что язык $\{xqx'\}$ не может быть порожден никакой КС-грамматикой. Строгое доказательство этого факта, см., например, Гладкий 1966, стр. 91.

Оказывается, что указанное явление встречается и в естественных языках, т. е. в них возможны фрагменты, состоящие из цепочек вида xqx'.

В литературе описаны два примера такого рода:

1) Конструкции типа:

a b c d a'
Миша, Перепетуя, Гриша, Фекла... — двоечник,
b' c' d'
отличница, троечник, четверочница ... соответственно.

Здесь роль x $(abcd...)$ играет цепочка собственных имен, а роль x' $(a'b'c'd'...)$ — цепочка характеристик, которые должны быть согласованы с этими именами в роде *); q — это тире (точнее, связка «*быть*» в нулевом выражении).

2) В индейском языке мохавк, как указывает П. Постал (Postal 1964), широко распространены предложения, в которых основное дополнение дублируется инкорпорированием соответствующих основ в глагол-сказуемое: *Девушка книго-читает книгу*. Кроме того, любой глагол (в том числе включающий инкорпорированные дополнения) легко субстантивируется и приобретает способность выступать в роли дополнения, в частности, инкорпорироваться в сказуемое: *Я книгочтение -интересуюсь книгочтением* (т. е. 'я интересуюсь чтением книги'). Этот процесс теоретически может быть повторен неограниченное число раз: *Вы книгочтениеинтересодумаете о книгочтениеинтересе* ('вы думаете об интересе к чтению книги'), *Они* $\overbrace{\textit{книго}}^{a}\overbrace{\textit{чтение}}^{b}\overbrace{\textit{интересо}}^{c}$$\overbrace{\textit{думание}}^{d}$ - $\overset{q}{\textit{видят}}$ $\overbrace{\textit{книго}}^{a'}\overbrace{\textit{чтение}}^{b'}\overbrace{\textit{интересо}}^{c'}\overbrace{\textit{думание}}^{d'}$ ('они видят думание об интересе к чтению книги') и т. д.

*) Аналогичный пример был впервые указан И. Бар-Хиллелом и Э. Шамиром (Bar-Hillel — Shamir 1960).

Здесь x' $(= a'b'c'd')$ — это дополнение, x $(= abcd)$ — его дубликат, инкорпорированный в сказуемое, а q — собственно сказуемое. Подчеркнем, что такая конструкция является правильной лишь тогда, когда инкорпорированный дубликат дополнения в точности соответствует самому дополнению по составу и порядку следования основ.

Если учитывать эти примеры, то приходится признать, что, вообще говоря, КС-грамматик недостаточно для описания любых естественных языков в полном объеме. Однако сразу видно, что оба примера носят периферийный характер: первая конструкция, хотя и допустима, вероятно, в любом языке, крайне специфична и не относится к числу употребительных, а вторая, имеющая очень общее значение и, по-видимому, достаточно употребительная, известна только в одном малораспространенном языке. Поэтому при всей теоретической ценности этих примеров ими можно пренебречь. Если же от них отвлечься, то КС-грамматики можно считать в принципе достаточным средством для описания естественных языков. Это утверждение, разумеется, не может быть строго доказано; убеждение в его истинности основывается на ряде следующих эмпирических соображений.

1) Существуют так называемые категориальные грамматики, относящиеся к распознающим грамматикам (см. стр. 122 и сл.). Эти грамматики разрабатывались и применялись к естественным языкам независимо от КС-грамматик, причем примеров их неадекватности (за исключением двух, указанных выше) до сих пор приведено не было. Однако доказано (см. ниже, стр. 133), что класс языков, описываемых категориальными грамматиками, совпадает с классом КС-языков.

2) Сравнительно недавно для описания языков были предложены автоматы с магазинной памятью *), способные осуществлять как распознавание, так и порождение. Н. Хомским было доказано (см. стр.

*) О них пойдет речь на стр. 136 и сл.

149), что все языки, обрабатываемые такими автоматами, суть КС-языки, и обратно. Таким образом, оказалось, что еще одна формальная модель естественного языка, введенная из независимых соображений и не встретившая существенных принципиальных трудностей, эквивалентна КС-грамматике.

3) В рамках того раздела математической лингвистики, который можно называть моделированием лингвистических исследований (см. ниже, стр. 162 и сл.), легко выделяется класс так называемых конечно-характеризуемых языков (см. стр. 170), которые интуитивно очень близки к естественным языкам. Оказывается, что все конечно-характеризуемые языки являются КС-языками (обратное неверно!). Это опять-таки склоняет к мысли о том, что КС-грамматики способны порождать естественные языки.

4) Наконец, известен целый ряд алгоритмов автоматического анализа и порождения текстов на естественных языках, которые (алгоритмы) используют в качестве описания соответствующих языков именно КС-грамматики или же эквивалентные им системы. Многие из этих алгоритмов запрограммированы для ЭЦВМ и опробованы в достаточно широких экспериментах; на КС-грамматиках основаны, например, алгоритмы синтаксического анализа для нескольких языков, разрабатываемые в Техасском университете (Tosh 1965), ряд алгоритмов, использующих так называемый метод Кока (Hays 1962, 412—414), и некоторые другие алгоритмы, упоминаемые в работе Bobrow 1963; ср. также Yngve 1961 и Арсентьева 1965.

Все это и заставляет признать КС-грамматики достаточными для естественных языков.

В частности, стоит отметить, что конструкции типа $abcd...d'c'b'a'$ (стр. 77), не описываемые А-грамматиками, легко могут быть порождены с помощью КС-грамматик. Так, легко показать, что язык, состоящий в точности из цепочек указанного вида (составленных из символов a_1, a_2, a_3, a_1', a_2', a_3'), порождается КС-грамматикой, содержащей всего

шесть правил:

$$\left.\begin{array}{l} I \to a_i I a_i' \\ I \to a_i a_i' \end{array}\right\} i = 1,\ 2,\ 3.$$

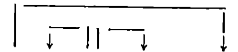
Теперь необходимо сделать два следующих важных замечания.

Во-первых, сказанное отнюдь не означает, будто КС-грамматики порождают только естественные языки или языки, близкие к ним: среди КС-языков имеются и такие, которые вовсе не похожи по своему строению на естественные.

Во-вторых, из того, что КС-грамматики практически достаточны для описания естественных языков, вовсе не следует, что они всегда удобны для этой цели, т. е. что они позволяют описывать любые конструкции естественных языков естественным образом. Более того, хорошо известно, что это не так (ср. аналогичное замечание относительно А-грамматик, стр. 79). КС-грамматики не обеспечивают, например, естественного («имеющего объяснительную силу») описания для так называемых непроективных конструкций *), т. е. для конструкций с разрывными составляющими (или, что то же самое, с пересечением ⸬, ⸬ и т. п. или обрамлением ⸬, ⸬ стрелок синтаксической зависимости). При этом непроективные конструкции изобилуют в самых разных языках:

Русск. *К этой поездке может пробудить интерес только выступление директора.*

*) О понятии проективности см. Падучева 1964б и Гладкий 1966, стр. 17 и сл.



Англ. *A theorem is stated which describes the properties of this function.*

Нем. ... *die Tatsache, daß die Menschen die Fähigkeit besitzen, Verhältnisse der objektiven Realität in Aussagen wiederzuspiegeln.*

Фр. ... *la guerre, dont la France portait encore les blessures...*

Сербо-хорв. *Regulacija procesa jedan je od najstarjih oblika regulacije.*

Венг. *Azt hisszem, hogy késedelmemmel sikerült bebizonyítani.*

Если мы хотим описывать строение подобных фраз в терминах составляющих (а грамматики, в частности КС-грамматики, описывают синтаксическую структуру именно так), то для естественного описания необходимо использовать разрывные составляющие: все слова, зависящие от одного и того же слова, должны образовывать (вместе с ним) одну составляющую, а это при отсутствии проективности обязательно поведет к появлению разрывных составляющих (*к этой поездке... интерес*, *a theorem ... which describes the properties of this function* и т. д.). Однако системы составляющих (НС-структуры), приписываемые фразам КС-грамматикой и, более того, любой НС-грамматикой, разрывных составляющих содержать не могут.

К этой трудности мы еще вернемся ниже, на стр. 106, а пока рассмотрим два специальных случая КС-грамматик, эквивалентных А-грамматикам (как это было обещано на стр. 80).

Первый случай. В естественных языках возможно размещение зависимых слов справа от главного («правое подчинение»): *лист бумаги*, *une règle stricte*, *give him*, или слева от главного («левое подчинение»): *белый лист*, *cette règle*, *good advice*.

Как правое, так и левое подчинение может быть последовательным: *жена сына заместителя председателя второй секции эклектики совета по прикладной мистике при президиуме Академии наук королевства Мурак* и *очень быстро бегущий олень*. В зависимости от языка та или иная конструкция с последовательным подчинением вправо или влево может быть теоретически неограниченной: такова, например, конструкция с последовательным подчинением родительных падежей в русском языке (неограниченное правое подчинение) и аналогичная конструкция в литовском (где $S_{\text{род}}$ всегда ставится перед подчиняющим его словом, что приводит к неограниченному левому подчинению). Тот факт, что языки мира различаются и могут классифицироваться по преобладанию в них правого или левого подчинения и, в частности, в зависимости от возможности неограниченного последовательного подчинения в ту или иную сторону, был отмечен и исследован еще Л. Теньером (Tesnière 1959, стр. 32—33). Недавно на эту проблематику—в связи с применением КС-грамматик для описания естественных языков — обратил внимание В. Ингве (Ингве 1965 и Yngve 1960). Он заметил, что существует большое количество языков (например, английский, русский, французский и т. п.), в которых последовательное правое подчинение принципиально не ограничено, а при левом подчинении длина цепочки всегда ограничена

в силу структурных особенностей этих языков *). Оказывается, что КС-грамматика, порождающая подобный язык, обладает следующим интересным свойством: для любой выводимой терминальной цепочки имеется такой вывод, в каждой строчке которого все вспомогательные символы собираются в правом конце, занимая не более чем K последних мест (K — константа, фиксированная для данной грамматики, т. е. одна и та же для всех выводов в ней)**).

Иначе говоря, если каждую строчку вывода разделить на две части: левую — одни терминальные символы до первого вспомогательного символа X — и правую — от X включительно до конца (в правой части могут содержаться и терминальные символы), то правая часть всегда будет содержать не более K символов. Левая часть содержательно интерпретируется как уже «выданный» кусок порождаемой цепочки (на следующих шагах вывода этот кусок больше не подвергается никакой переработке), а правая — как рабочий участок, который грамматика должна, так сказать, держать в памяти. Таким образом, число K есть не что иное, как максимальный объем памяти, необходимый для порождения любой цепочки в данной грамматике (т. е. найдется цепочка, не порождаемая при объеме памяти $< K$). Это число совпадает с максимальной длиной цепочки последовательных левых подчинений, возможной в рассма-

*) Мы не будем останавливаться здесь на так называемой «гипотезе Ингве» (Ингве 1965), которая представляет собой попытку объяснить это эмпирическое наблюдение некоторыми общими закономерностями строения человеческой психики.

**) Строго говоря, для того чтобы КС-грамматика обладала указанным свойством, ограниченности последовательного левого подчинения недостаточно. Необходимо выполнение ряда более сильных и сложно формулируемых требований (см. Падучева 1967), из которых вытекает, например, ограниченность правого параллельного подчинения ...

и последовательного вложения типа ... По этому вопросу см. также Шрейдер 1966.

триваемом языке: так, если в каком-то языке не бывает более трех последовательных подчинений влево, то при порождении этого языка для любой цепочки можно построить такой вывод, в котором не возникает необходимости запоминать более трех символов сразу. Отмеченная связь между допустимой глубиной левого подчинения и объемом памяти была установлена В. Ингве (там же).

Проиллюстрируем сказанное примером, а именно рассмотрим грамматику Γ_4, порождающую некоторые именные группы русского языка, в которых правое подчинение не ограничено, а глубина левого не превосходит двух.

Схема грамматики Γ_4

$\widetilde{S}_{x,\ y,\ z} \rightarrow S_{x,y,z}\widetilde{S}_{x',\ y',\ \text{род}}$

$\widetilde{S}_{x,\ y,\ z} \rightarrow \widetilde{A}_{x,y,z}\widetilde{S}_{x,\ y,\ z}$

$\widetilde{A}_{x,\ y,\ z} \rightarrow \left\{\begin{matrix} \textit{очень, совершенно,} \\ \textit{весьма,} \ldots \end{matrix}\right\} A_{x,\ y,\ z}$

$\widetilde{S}_{x,\ y,\ z} \rightarrow S_{x,\ y,\ z}$

$\widetilde{A}_{x,\ y,\ z} \rightarrow A_{x,\ y,\ z}$

$S_{\text{м},\ y,\ z} \rightarrow \textit{сын}_{y,\ z},\ \textit{заместитель}_{y,\ z},\ \textit{председатель}_{y,\ z},\ \textit{комитет}_{y,\ z}, \ldots$

$S_{\text{ж},\ y,\ z} \rightarrow \textit{жена}_{y,\ z},\ \textit{секция}_{y,\ z}, \ldots$

$A_{x,\ y,\ z} \rightarrow \textit{красивый}_{x,\ y,\ z},\ \textit{толстый}_{x,\ y,\ z},\ \textit{важный}_{x,\ y,\ z}, \ldots$

(Смысл обозначений — см. стр. 55 и 57; три последние строчки представляют собой сокращенную запись, смысл которой очевиден. Как и в Γ_1 и Γ_3, здесь не учтены особенности согласования A с одушевленным $S_{\text{вин}}$.) Приведем пример вывода в грамматике Γ_4:

$\widetilde{S}_{\text{ж, ед, вин}}$

$\widetilde{A}_{\text{ж, ед, вин}}\widetilde{S}_{\text{ж, ед, вин}}$

очень $A_{\text{ж, ед, вин}}\widetilde{S}_{\text{ж, ед, вин}}$

очень красивую $\widetilde{S}_{\text{ж, ед, вин}}$

очень красивую $\widetilde{S}_{ж,\ ед,\ вин}\widetilde{S}_{м,\ ед,\ род}$

очень красивую $S_{ж,\ ед,\ вин}\widetilde{S}_{м,\ ед,\ род}$

очень красивую жену $\widetilde{S}_{м,\ ед,\ род}$

очень красивую жену $\widetilde{S}_{м,\ ед,\ род}\widetilde{S}_{м,\ ед,\ род}$

очень красивую жену $S_{м,\ ед,\ род}\widetilde{S}_{м,\ ед,\ род}$

очень красивую жену сына $\widetilde{S}_{м,\ ед,\ род}$

. . . .

очень красивую жену сына заместителя председателя $S_{ж,\ ед,\ род}$

очень красивую жену сына заместителя председателя секции

В этом выводе объем памяти равен двум: ни одна промежуточная цепочка не содержит более двух вспомогательных символов. Ту же цепочку можно было бы породить и по-другому, используя больший объем памяти, например, сначала получить из $\widetilde{S}_{ж,ед,вин}$ цепочку

$$очень\ A_{ж,\ ед,\ вин}S_{ж,\ ед,\ вин}S_{м,\ ед,\ род}S_{м,\ ед,\ род}\ S_{м,\ ед,\ род}\ S_{ж,\ ед,\ род},$$



а уже из нее нашу терминальную цепочку. Для нас, однако, важен необходимый объем памяти, т. е. такой, что с меньшим объемом получить данную цепочку невозможно. Именно этот объем и равен здесь двум.



Можно доказать, что и любая терминальная цепочка, выводимая в Γ_4, может быть порождена с объемом памяти $\leqslant 2$. Доказательство (опущенное здесь) основано на очень простом соображении: «хороший» вывод надо проводить так, чтобы для каждого существительного сначала выдавались в терминальном виде его левые зависимые, и только потом выполнялось развертывание именной группы вправо.

$\mathrm{T}_{1.1.6}$ КС-грамматика описанного типа («КС-грамматика с ограниченной памятью») всегда эквивалентна некоторой автоматной грамматике. Это нетрудно до-

казать (идея доказательства, которое мы здесь не приводим, заключается в том, что правая часть строчки вывода, состоящая из K символов, кодируется одним новым вспомогательным символом). Таким образом, мы видим, что в случае языков с ограниченной глубиной левого подчинения КС-грамматики с ограниченной памятью, эквивалентные А-грамматикам и близкие к ним по построению выводов, т. е. устроенные гораздо проще, чем произвольные КС-грамматики, оказываются не только принципиально достаточными, но и весьма удобными — они обеспечивают достаточно естественное описание.

Второй случай. Бывают, однако, языки, в которых неограниченную глубину имеет не только правое, но и левое последовательное подчинение. Подобным языком является, например, венгерский, где неограниченное левое подчинение получается за счет препозитивных распространенных определений *), а неограниченное правое подчинение — за счет, например, придаточных предложений с *который* (ср. *Дом, который построил Джек*).

---

*) См. пример из новеллы Г. Фехера — шуточный тост, приведенный в работе Varga 1964, стр. 70:

Kivánom, hogy valamint az agyag [23] *ölelő karjai* [22] *közül kibontakozni* [21] *akaró* [20] *kocsikerék* [19] *rettentő nyikorgásától* [18] *megriadt* [17] *juhászkutya* [16] *bundájába* [15] *kapaszkodó* [14] *kullancs* [13] *kidülledt félszeméből* [12] *alácseppent* [11] *könnycseppben* [10] *visszatükröződő* [9] *holdvilág fényétől* [8] *illuminált* [7] *rablólovagvár* [6] *felvonóhidjából* [5] *kiálló* [4] *vasszegek* [3] *kohéziós erejének* [2] *hatása* [1] *évszázadokra összetartja annak materiáját, aképpen tartsa össze ezt a társaságot az igaz szeretet.*—

Я хочу, чтобы настоящая любовь скрепила эту компанию так, как на столетия скрепляет материал моста действие [1] *связующей силы* [2] *гвоздей* [3], *торчащих* [4] *из подъемного моста* [5] *разбойничьего феодального замка* [6], *освещенного* [7] *лунным светом* [8], *отражающимся* [9] *в слезинке* [10], *вытекшей* [11] *из выпученного глаза* [12] *клеща* [13], *вцепившегося* [14] *в шерсть* [15] *овчарки* [16], *встревоженной* [17] *ужасным скрипом* [18] *тележных колес* [19], *жаждущих* [20] *вырваться*.[21] *из объятий* [22] *грязи* [23]. Эта фраза— не придуманная, а взятая из художественного текста! — имеет глубину 22 и является абсолютно правильной с грамматической точки зрения (точно в такой же степени, как и ее русский перевод). Более того, ничто не мешает продолжить цепь определений влево ad libitum.

Для порождения языков с таким свойством можно предложить еще один особый тип КС-грамматик, в некотором смысле более общий, чем КС-грамматики с ограниченной памятью, рассмотренные выше.

Прежде всего сформулируем более точно, какие языки мы имеем здесь в виду. Это языки, в которых возможно неограниченное число последовательно подчиненных слева направо конструкций $X_1X_2\ldots$ $\ldots X_i\ldots$ (неограниченное правое подчинение), и при этом в каждой из конструкций X_i возможно неограниченное левое подчинение — последовательность конструкций$\ldots X_{ij}\ldots X_{i2}X_{i1}$; однако внутри конструкций X_{ij} дальнейшее неограниченное развертывание невозможно. Применительно к венгерскому языку X_i можно понимать как простые предложения, являющиеся каждое (кроме первого) придаточным определительным к предыдущему, а X_{ij} — как препозитивные причастные обороты (ср. пример в сноске на стр. 101).

Представим себе грамматику

$$\Gamma^* = \langle \boldsymbol{V}^*,\ \boldsymbol{V}_1^*,\ \boldsymbol{I}^*,\ \boldsymbol{S}^* \rangle,$$

основной словарь которой $\boldsymbol{V}^*$ состоит из n символов $A_1, A_2, \ldots, A_n$ и правила которой имеют вид

$$X \to YA_i \text{ или } X \to A_i,$$

где X и Y принадлежат к $\boldsymbol{V}_1^*$. Поставим в соответствие каждому из символов A_i некоторую А-грамматику $\Gamma_i^* = \langle \boldsymbol{V},\ \boldsymbol{V}_1^i,\ A_i,\ \boldsymbol{S}_i \rangle$ где $\boldsymbol{V}$ — основной словарь, общий для всех Γ_i^*, $\boldsymbol{V}_1^i$ — вспомогательный словарь, не содержащий никаких символов из $\boldsymbol{V}^*$ и $\boldsymbol{V}_1^*$, кроме A_i; A_i — начальный символ; правила схемы $\boldsymbol{S}_i$ имеют вид $C \to dD$ или $C \to c$ (здесь, как и в других примерах, заглавными латинскими буквами обозначаются вспомогательные символы, а строчными — основные). При этом будем считать, что вспомогательные словари грамматик Γ_i^* попарно не пересекаются.

Грамматика Γ^* очень близка к автоматной, отличаясь от нее только направлением развертыва-

ния*) порождаемой цепочки; в сущности она является автоматной с точностью до зеркальной симметрии. Таким образом, мы имеем дело с одной «праворазвертывающей» (квази-А-) грамматикой и с n «леворазвертывающими» А-грамматиками.

Рассмотрим теперь объединение всех этих грамматик, точнее, грамматику Г, у которой основной словарь — $\boldsymbol{V}$ (тот же, что у всех Γ_i^*), вспомогательный словарь — $\boldsymbol{V}_1 = \boldsymbol{V}^* \cup \overset{*}{\boldsymbol{V}_1} \cup \boldsymbol{V}_1^1 \cup \boldsymbol{V}_1^2 \cup \ldots$ $\ldots \cup \boldsymbol{V}_1^n$ (т. е. объединение вспомогательных словарей всех грамматик Γ^*, $\Gamma_1^*, \ldots, \Gamma_n^*$ и основного словаря грамматики Γ^*), начальный символ — $\boldsymbol{I}$ (тот же, что у Γ^*), а схема есть объединение схем всех грамматик Γ^*, Γ_1^*, $\Gamma_2^*, \ldots$, Γ_n^*. Эта грамматика Г представляет собой специальную КС-грамматику, которую можно назвать К С-г р а м м а т и к о й с н е з а в и с и м ы м д в у с т о р о н н и м р а з в е р т ы в а н и е м **).

Приведем пример (схемы) такой грамматики.

$$\boldsymbol{S}^* = \begin{cases} I \to BA_1 \\ B \to CA_1 \\ C \to BA_2 \\ C \to DA_3 \\ D \to DA_4 \\ D \to A_2 \end{cases} \qquad \boldsymbol{S}_3 = \begin{cases} A_3 \to aP_3 \\ A_3 \to bQ_3 \\ A_3 \to cR_3 \\ P_3 \to a \\ Q_3 \to b \\ R_3 \to dR_3 \\ R_3 \to eR_3 \\ R_3 \to d \end{cases}$$

$$\boldsymbol{S}_1 = \begin{cases} A_1 \to bP_1 \\ P_1 \to aQ_1 \\ Q_1 \to aQ_1 \\ Q_1 \to c \end{cases} \qquad \boldsymbol{S}_4 = \begin{cases} A_4 \to cP_4 \\ P_4 \to b \end{cases}$$

$$\boldsymbol{S}_2 = \{ \ A_2 \to d$$

---

*) Под направлением развертывания здесь понимается направление, в котором «выбрасываются» терминальные символы, например, $C \to dD$ — левое развертывание.

**) То, что эта грамматика не является автоматной, очевидно хотя бы потому, что некоторые ее правила (правила

Грамматика введенного нами типа работает следующим образом. Сначала порождаемая цепочка неограниченно развертывается слева направо за счет символов A_i (которые могут интерпретироваться, например, как синтаксические группы или предложения); это делается правилами $\boldsymbol{S}^*$. Затем любое из $\boldsymbol{A}_i$ может (правилами $\boldsymbol{S}_i$) неограниченно развертываться справа налево — в цепочку терминальных символов (которые можно интерпретировать как слова). Подобный процесс порождения удобен в таких, например, случаях, как венгерские фразы рассмотренного выше типа.

Каждая КС-грамматика с независимым двусторонним развертыванием эквивалентна некоторой А-грамматике. Доказательство этого факта мы приводить не будем. **Т**1.1.7

Использование НС-грамматик и неукорачивающих грамматик для описания естественных языков

Как отмечалось выше, НС-грамматики (и равные им по порождающей силе неукорачивающие грамматики, см. стр. 54) представляют собой лишь частный случай общего понятия грамматики. Тем не менее, НС-грамматики безусловно достаточны (хотя и не обязательно удобны) для описания любых естественных языков в полном объеме. Это вытекает из следующего практически очевидного допущения: любой естественный язык (точнее, множество его правильных фраз) есть легко распознаваемое множество. Напомним, что это означает (см. стр. 50) существование достаточно простого алгоритма распознавания правильности фраз; несомненно, что говорящие обладают таким алгоритмом. Разумно было

схемы S) содержат в правых частях по два вспомогательных символа. В самом деле, основные символы грамматики Γ^* (т. е. $A_1, A_2, \ldots, A_n$) в грамматике Γ являются вспомогательными, так что правила вида $X \to YA_i$ в рамках Γ оказываются не «автоматными». Однако грамматика Γ эквивалентна автоматной, см. ниже.

бы, по-видимому, полагать, что алгоритм распознавания правильности фраз естественного языка должен обеспечивать такой процесс распознавания, при котором требуемый объем «оперативной памяти» сопоставим с длиной фразы — например, не превышает числа Mn, где n — длина фразы, а M — достаточно большое число, не зависящее от n (т. е. некоторая константа). Это допущение тем более естественно, что, как известно из психологических экспериментов, объем оперативной памяти человека вообще ограничен и притом весьма невелик *).

Если же язык распознается алгоритмом с указанным ограничением на объем памяти («алгоритмом с ограниченным растяжением»), то он может быть порожден грамматикой, в которой для любой выводимой терминальной цепочки длины n существует такой вывод, в котором ни одна промежуточная цепочка не превосходит по длине числа Kn (K — некоторая константа)**). Такую грамматику можно назвать грамматикой с ограниченным растяжением; более точно, **грамматика с ограниченным растяжением** — это грамматика, у которой емкостная сигнализирующая функция (стр. 119) не больше линейной.

Доказано, что для любой грамматики с ограниченным растяжением может быть построена эквивалентная ей НС-грамматика. Таким образом, если принять оба наши допущения (а они, по-видимому, должны быть приняты), то приходится признать, что НС-грамматики в принципе способны описывать множество правильных фраз любого естественного языка, т. е. порождать любые правильные фразы данного языка, не порождая при этом никаких неправильных. (В частности, обе конструкции,

$T_{1.3.4}$

---

*) С фактом ограниченности объема оперативной памяти человека связаны и те лингвистические соображения, которые привели к понятию «грамматики с ограниченной памятью», см. выше, стр. 97—101.

**) Это утверждение может быть строго сформулировано и доказано, если надлежащим образом уточнить понятие алгоритма, например, через понятие машины Тьюринга.

приведенные выше как примеры «непригодности» КС-грамматик, легко описываются НС-грамматиками. В самом деле, грамматика на стр. 91, порождающая язык $\{xqx'\}$, как отмечалось во второй сноске на той же странице, может быть заменена эквивалентной НС-грамматикой.)

Но это лишь одна сторона дела. Другая сторона, как и в предыдущих случаях,—это удобство и естественность описания. Известно, что проблема приемлемости и границ «анализа по непосредственно составляющим» неоднократно дискутировалась в лингвистике. Не рассматривая эту сложную и интересную проблему во всех подробностях, мы попытаемся резюмировать ее основное содержание. Отмечаемые обычно недостатки метода НС, т. е. фактически НС-грамматик, сводятся к трем пунктам.

1) С помощью НС-грамматик не удается естественно описывать фразы, содержащие разрывные составляющие. Об этом уже говорилось выше, применительно к КС-грамматикам (стр. 95). Поскольку любые НС-грамматики сопоставляют фразе систему составляющих в точности таким же образом, как и КС-грамматики, то все сказанное выше о непроективных фразах относится и к НС-грамматикам.

Пункты 2) и 3) касаются в сущности не только НС-грамматик, но и вообще всех порождающих грамматик в строгом смысле определения на стр. 48.

2) НС-грамматика, как и любая порождающая грамматика, содержит только правила образования языковых выражений, например словоформ или фраз. Это значит, что грамматика задает правильные выражения в отличие от неправильных. Сам термин «правила образования» заимствован из математической логики, где он обозначает правила построения правильных формул *).

*) Например, алгебраическая формула $(a + b)\,c$ является правильной, а формула $(a +)$ — неправильной.

Однако в логике рассматривается еще и другой тип правил — п р а в и л а п р е о б р а з о в а н и я. Эти последние задают определенные соотношения между правильными формулами, причем по существу это смысловые отношения. Так, в элементарной алгебре это (известные всем из средней школы) правила тождественных преобразований: из одного выражения получается другое с тем же числовым значением, которое (значение) играет здесь роль смысла. В алгебре логики это правила вывода, позволяющие из одних истинных выражений получать другие истинные выражения; здесь в качестве смысла выступает «истинностное значение», т. е. истинность или ложность. Введение правил преобразования всегда означает переход к более высокому уровню рассмотрения языка, а именно к семантическому уровню.

Совершенно очевидно, что правила преобразования необходимы и при описании естественных языков (ср. стр. 153—154). Владение языком обязательно предполагает умение не только построить правильную фразу, но и перейти от одной фразы к другим, либо полностью синонимичным ей, либо отличающимся от нее по смыслу на определенную «величину». Так, говорящий легко делает из утвердительного предложения вопросительное или отрицательное, из активной конструкции пассивную, без труда меняет стилистическую окраску текста, выражает одну и ту же мысль разными способами и т. д. Эти возможности, которые обязательно должны учитываться в описании языка, не могут быть изложены в терминах грамматик, и поэтому встает вопрос о разработке формального аппарата для правил преобразования применительно к естественным языкам.

Соответствующая задача была впервые четко сформулирована в работах Н. Хомского (Хомский 1962). Выдвинутая им концепция быстро приобрела широкую популярность под именем «трансформационной грамматики». (Сам термин «трансформация» — по-английски transformation — в действительности

означает 'преобразование', так что rules of transformations = правила преобразования.) Пафос этой концепции, по нашему мнению, состоит во введении еще одного — семантического—уровня описания языка*). В самом деле, инвариантом всех трансформаций обычно является смысл; иначе говоря, трансформации — это преобразования, сохраняющие (или почти сохраняющие) смысл. Таким образом, теория трансформаций оказывается по существу теорией синонимии в языке (имеется в виду синонимия в самом широком смысле слова, прежде всего синонимия предложений, а также более крупных отрезков текста **)). В последнее время становится все более очевидным, что описание синонимии должно занимать в лингвистике одно из центральных мест. Отсюда вытекает и первостепенная роль трансформаций.

Однако трансформации относятся не к тому же уровню, что НС-грамматики: НС-грамматики относятся к синтаксическому (в широком смысле слова, см. ниже, стр. 153) уровню, а трансформации— к семантическому. Поэтому, когда говорят о недостаточности НС-грамматик для описания языка, следует помнить, что это верно только в смысле неохвата НС-грамматиками семантического уровня; на своем, чисто синтаксическом уровне НС-грамматики оказываются принципиально вполне достаточными.

Порождающие грамматики (в смысле определения на стр. 48) рассматриваются в рамках строго формальной теории; что же касается трансформаций, то здесь подобный уровень формализации пока не достигнут: предлагавшиеся до сих пор трансформационные правила не сформулированы в терминах одной простой операции, как правила грамматик (с их операцией подстановки). Задача дальнейшей

---

*) Подробнее об описании семантического уровня языка см. ниже, стр. 153 и сл.

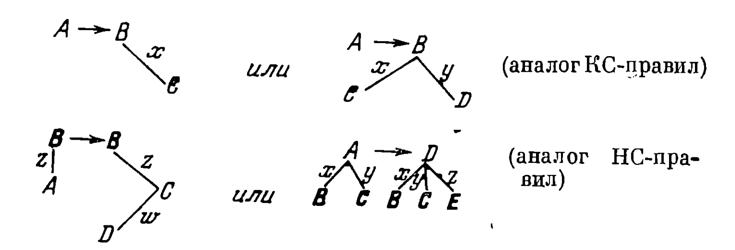
**) В работах З. Хэрриса (например, Хэррис 1962) термин «трансформация» имеет другое значение.

формализации трансформаций является весьма актуальной.

В заключение сделаем следующее замечание терминологического характера. В работах Н. Хомского и некоторых других авторов термин «порождающая грамматика» используется в двух смыслах: в широком — для обозначения любой системы формальных правил, описывающих язык, с включением трансформационного и морфонологического компонентов, и в узком — для обозначения грамматик в смысле определения на стр. 48. В нашем изложении этот термин всегда используется только в узком смысле. При подобном словоупотреблении трансформационные правила оказываются за рамками порождающей грамматики.

3) НС-грамматика, как и любая порождающая грамматика, *строит предложения сразу с точно определенным порядком слов* — с тем, который эти предложения должны иметь в окончательном виде. При этом порождаемому предложению сопоставляется синтаксическая структура в форме *упорядоченного дерева*, т. е. дерева, где между узлами, кроме отношения подчинения, задаваемого самим деревом, имеется еще и отношение линейного порядка (правее — левее). Таким образом, в синтаксической структуре, определяемой НС-грамматиками, не расчленены два совершенно различных по своей природе, хотя и связанных между собой отношения: синтаксическое подчинение и линейное взаиморасположение. Между тем в лингвистической традиции всегда с полным основанием считалось, что охарактеризовать синтаксическую структуру — это значит указать отношения синтаксического подчинения (как бы ни толковалось это последнее понятие). Что же касается отношения линейного порядка, то оно должно характеризовать не структуру, а саму фразу. Порядок слов, конечно, зависит от синтаксической структуры; он определяется обязательно с ее учетом и тем самым является по отношению к ней чем-то производным, вторичным.

Поскольку это так, представляется целесообразным видоизменить понятие порождающей грамматики таким образом, чтобы левые и правые части правил подстановки представляли собой не линейно упорядоченные цепочки, а например, деревья (без линейной упорядоченности!), изображающие синтаксические отношения. Тогда правила могли бы иметь, скажем, такой вид:



Черточки с индексами изображают синтаксические связи различных типов; буквы A, B, C,...— синтаксические категории. NB: взаимное расположение символов одного уровня подчинения не играет никакой роли и является на данной схеме случайным: означает в точности то же самое, что и .

В результате мы получили бы исчисление синтаксических структур (а не фраз!) языка. Это исчисление не было бы порождающей грамматикой в узком смысле этого термина, однако его вполне можно рассматривать как часть порождающей грамматики в широком смысле; необходимо, однако, построить соответствующее формальное определение. Другую часть этой грамматики составляло бы исчисление, которое для любой данной синтаксической структуры задавало бы (возможно, с учетом каких-либо

других факторов, например, в русском языке — с обязательным учетом логического выделения и т. п.) все возможные для нее линейные последовательности слов. При таком подходе, в частности, снимается проблема разрывных составляющих, см. выше, стр. 106. Пример исчисления, порождающего синтаксические структуры, а затем выполняющего линейное упорядочение их терминальных символов, дан в работе Ломковская 1965—1966; вариант системы правил упорядочения слов в русском языке (исходя из синтаксической структуры, свойств самих слов и сведений о логическом выделении) содержится в Мельчук 1965; исчисление синонимии русских фраз («система перифразирования»), в котором формирование структур фраз отделено от построения соответствующих последовательностей лов, описано в Жолковский — Мельчук 1967.

§ 5. *О формальных свойствах порождающих грамматик*

В предыдущем параграфе мы коснулись ряда вопросов, связанных с применением формальных порождающих грамматик для описания естественных языков. Там мы рассматривали в основном те свойства грамматик, которые представляют непосредственный интерес именно в плане лингвистических приложений, т. е. исследовали грамматики с точки зрения содержательной интерпретации. Теперь мы остановимся на некоторых «внутренних», чисто формальных свойствах грамматик, т. е. сообщим отдельные факты из математической теории порождающих грамматик. Грамматики представляют собой чисто математический объект и как таковой могут изучаться чисто математическими средствами. Далеко не все результаты такого изучения имеют лингвистическую интерпретацию; многие из них сами по себе вряд ли нужны для исследования естественных языков. Тем не менее все факты, относящиеся к грамматикам,— мы имеем в виду их формальные свойства — занимают в единой математической теории грамматик определенное место и в рамках этой теории необходимы. Дело в том, что любая теория развивается по своим внутренним законам и может существовать только как целое. Выделить в ней некие особые разделы, интересные для приложений, и развивать их отдельно невозможно. Поэтому в той степени, в какой сама теория грамматик в целом признается полезной и важной для лингвистики, в конечном счете (т. е. хотя бы и косвенно) оказываются нужными и чисто математические результаты.

Математическая теория грамматик в настоящее время довольно хорошо разработана. Мы остановимся здесь только на двух ее разделах — на алгоритмических проблемах и на проблеме оценки сложности вывода.

Алгоритмические проблемы

В теории грамматик часто возникают вопросы о наличии/отсутствии того или иного алгоритма, например:

1) Существует ли алгоритм, позволяющий по любой данной КС-грамматике узнать, является ли порождаемый ею язык конечным (т. е. конечно ли порождаемое ею множество терминальных цепочек)?

2) Аналогичный вопрос можно задать и относительно НС-грамматик.

3) Существует ли для данной грамматики Г алгоритм, позволяющий по паре (терминальных) цепочек x и y, выводимых в данной грамматике, определить, замещаемо *) ли x на y в языке L(Г)?

4) Существует ли алгоритм, позволяющий узнать относительно любой КС-грамматики, приписывает ли она каждой порождаемой терминальной цепочке только одну синтаксическую структуру, т. е. имеются ли терминальные цепочки, которые могут быть выведены в данной грамматике более чем одним способом?

Подобные вопросы и образуют круг алгоритмических проблем теории грамматик.

Сформулируем некоторые результаты, относящиеся к данной области.

$\mathbf{T}_{1.2.1}$ 1. В классе всех грамматик ни одно нетривиальное свойство языков, порождаемых грамматиками, не распознаваемо. (Мы называем свойство языков нетривиальным в данном классе языков, если в этом классе имеются как языки, обладающие этим свойством, так и языки, им не обладающие. Например, свойство «порождаться грамматикой» в классе языков, порождаемых грамматиками, тривиально, а

*) Определение замещаемости см. на стр. 166.

свойство «быть конечным» не тривиально.) Нераспознаваемость нетривиальных свойств означает, что если иметь дело с грамматиками, на которые не наложено никаких ограничений, то ни для какого нетривиального свойства языков не существует алгоритма, позволяющего по любой грамматике узнать, обладает ли порождаемый ею язык этим свойством.

2. В классе НС-грамматик:

$T_{1.2.2}$ а) Свойство порождать язык, содержащий данную цепочку, распознаваемо, т. е. для каждой цепочки существует алгоритм, позволяющий по любой НС-грамматике узнать, входит ли эта цепочка в порождаемый данной грамматикой язык. Доказательство этого факта по существу содержится на стр. 50—53, где был указан алгоритм, позволяющий по произвольной неукорачивающей грамматике Г и произвольной цепочке x узнать, выводима ли x в Г (т. е. принадлежит ли x к L(Г)). Важно отметить, что алгоритм этот один и тот же для всех Г и всех x и, таким образом, ни от Г, ни от x не зависит. Стало быть, распознаваемость свойства содержать данную цепочку доказана для класса неукорачивающих грамматик и тем более для его подкласса — НС-грамматик.

Отсюда следует распознаваемость также и некоторых сходных свойств—таких, как содержать одновременно 2,3,..., n данных цепочек, содержать хотя бы одну из n данных цепочек, содержать данную цепочку x при условии, что другая данная цепочка y (определенным образом связанная с x) принадлежит языку и т. п.

$T_{1.2.3}$ б) Практически все остальные «хорошие» (естественно возникающие) нетривиальные свойства языков нераспознаваемы. Так, для НС-грамматик нераспознаваемы свойства порождать данный (произвольный) НС-язык, порождать конечный язык, порождать КС-язык, порождать язык с замещаемостью x на y (где x и y — произвольные фиксированные цепочки) и т. д.

3. В классе КС-грамматик оказываются распознаваемыми некоторые из тех свойств, которые не-

распознаваемы в классе НС-грамматик. (Разумеется, свойства, распознаваемые для НС-грамматик (ср. пункт 2а), распознаваемы и для КС-грамматик.) Так, здесь распознаваемы следующие свойства:

а) Свойство порождать пустой язык (существует алгоритм, позволяющий для любой КС-грамматики узнать, порождает ли она хотя бы одну терминальную цепочку). $T_{1.2.4}$

б) Свойство порождать конечный язык.

в) Свойство порождать хотя бы одну цепочку, содержащую вхождение данной цепочки x. В лингвистической интерпретации это может означать, например, следующее: по КС-грамматике языка для любого словосочетания можно определить, входит ли оно хотя бы в одну фразу языка. Таким образом, в отличие от НС-грамматик, где распознаваема правильность целых фраз, но не частей фраз, для КС-грамматик распознаваемо и то и другое.

Тем не менее многие важные свойства нераспознаваемы и для КС-грамматик. В частности, нераспознаваемы следующие свойства: порождать А-язык; порождать полный язык (содержащий всевозможные цепочки, составленные из терминальных символов); иметь эквивалентную КС-грамматику, приписывающую каждой терминальной цепочке только одну синтаксическую структуру, и т. д. $T_{1.2.5}$

Заметим еще, что для КС-грамматик неразрешима проблема распознавания эквивалентности (произвольных двух грамматик), т. е. не существует алгоритма, позволяющего по любой паре КС-грамматик узнать, являются ли они эквивалентными. $T_{1.2.7}$

4. В классе А-грамматик распознаваемы практически все «интересные» свойства, в том числе все перечисленные в предыдущих пунктах; проблема распознавания эквивалентности здесь также разрешима. $T_{1.2.6}$ $T_{1.2.8}$

Из проблем, относящихся к алгоритмам не для классов грамматик, а для конкретных фиксированных грамматик, мы упомянем только одну: так называемую проблему распознавания замещаемости. Она состоит в том, чтобы для данной грамматики Г

найти алгоритм, позволяющий по любой паре цепочек x, y узнать, замещаема ли x на y в языке $L(\Gamma)$. Для некоторых грамматик такие ал- $T_{1.2.10}$ горитмы существуют, в частности, для всех А-грамматик. Однако имеются примеры КС-грамматик, $T_{1.2.9}$ для которых подобного алгоритма нет.

Оценки сложности вывода

Кроме алгоритмических проблем, к математической теории грамматик относятся также проблемы оценки сложности вывода в грамматиках. Сложность вывода естественно измерять либо числом шагов, т. е. числом промежуточных цепочек, либо необходимым объемом «памяти», например длиной промежуточных цепочек *). Для оценки сложности по числу шагов предложена мера

$$\tau_\Gamma(n) = \max_{l(x) \leqslant n} \tau_{\Gamma, x},$$

где Γ — данная грамматика; x — произвольная цепочка, выводимая в Γ; $l(x)$ — длина цепочки x (число символов в x); $\tau_{\Gamma,x}$—число шагов («время») самого короткого вывода цепочки x в Γ; n — произвольное натуральное число. Чтобы найти $\tau_\Gamma(n)$, нужно, как видно из формулы, найти для каждой выводимой цепочки, по длине не превосходящей n, наименьшее «время» (число шагов) вывода этой цепочки в Γ (дело в том, что цепочка x может иметь в Γ много разных выводов разной длины); затем среди всех этих «времен» берется максимальное. Это и будет значением функции $\tau_\Gamma(n)$ — так называемой в р е м е н н о́ й с и г н а л и з и р у ю щ е й ф у н к ц и и. Другими словами, $\tau_\Gamma(n)$—это такое число шагов, которое, с одной стороны, заведомо достаточно для вывода любой цепочки не длиннее n, а с другой стороны, необходимо: среди выводимых цепочек не длиннее n имеется хотя бы одна цепочка, которую нельзя вывести меньше чем в $\tau_\Gamma(n)$ шагов.

*) Объем памяти может измеряться и иначе: числом вспомогательных символов в промежуточных цепочках, расстоянием от первого вспомогательного символа до конца цепочки (ср. стр. 98) и т. п.

Исходя из понятия временно́й сигнализирующей функции, можно получать оценки сложности вывода в разных грамматиках. Так, на стр. 51—52 фактически было показано, что для любой неукорачивающей грамматики Г имеет место неравенство $\tau_\Gamma(n) < p^{n+1}$, где p — общее число основных и вспомогательных символов. Очевидно, что это неравенство имеет силу, в частности, и для НС-, и для КС-грамматик. Однако для последних указанную оценку можно значительно улучшить (т. е. понизить): для любой КС-грамматики Γ_1 выполняется неравенство $\tau_\Gamma(n) \leqslant \leqslant 2Cn$, где C—число вспомогательных символов в Γ_1. Эта оценка получается из следующих несложных соображений: любой КС-вывод можно провести так, чтобы на каждом шаге заменялся самый левый вспомогательный символ; на каждом шаге длина цепочки либо не изменяется, либо увеличивается, причем возможны шаги трех типов:

T$_{1.3.1}$

T$_{1.3.2}$

1. $A \to B$ («неудлиняющий нетерминальный», A и B — вспомогательные символы);

2. $A \to a$ («неудлиняющий терминальный», A — вспомогательный, a — основной символ);

3. $A \to d_1 \ d_2 \dots \ d_m$, $m > 1$ («удлиняющий»).

Если вывод не содержит «петель» (повторений промежуточных цепочек; ср. стр. 52), то в нем нигде не может быть больше чем C «неудлиняющих нетерминальных» шагов подряд. В самом деле, если бы было произведено подряд k таких шагов, где $k \geqslant C$, то соответствующие цепочки вывода имели бы вид:

xA_0Y

1-й шаг

xA_1Y

2-й шаг

xA_2Y

. . . .

C-й шаг

xA_CY

. . . .

k-й шаг

xA_kY

Здесь x — цепочка основных символов; A_0, A_1,... ...,A_k — вспомогательные символы; Y — некоторая цепочка произвольного вида. Уже C шагов дают $(C + 1)$ цепочек; поскольку эти цепочки различаются только символами A_0, A_1,..., A_C, а среди этих символов — не более C различных, две из этих цепочек обязательно совпадут, т. е. получится «петля». Таким образом, в любом КС-выводе без «петель», а тем более в кратчайшем выводе не бывает более $(C - 1)$ неудлиняющих нетерминальных шагов подряд. Между двумя сериями таких шагов обязательно должен вклиниваться хотя бы один шаг типа 2 или 3; а таких шагов всего не более $2n$ (не более n терминальных, поскольку вся цепочка состоит из n терминальных символов, и не более n «удлиняющих», поскольку каждый «удлиняющий» шаг увеличивает длину цепочки, а увеличиться больше чем на n эта длина не может). Стало быть, мы имеем не больше чем $2n$ серий шагов типа 1 (не более чем по $(C - 1)$ шагов в серии) и не более чем $2n$ шагов типа 2 и 3; общее же число шагов не превосходит $2n(C - 1) + 2n = 2Cn$, что и требовалось доказать.

Мы проиллюстрировали получение верхних оценок сигнализирующей функции. Получение нижних оценок оказывается более сложным. Приведем без доказательства одну такую оценку; отметим, что другие нижние оценки пока неизвестны.

$T_{1.3.3}$ Язык, состоящий из всевозможных цепочек вида xqx' (см. стр. 91), не может быть порожден никакой такой НС-грамматикой Г, у которой временна́я сигнализирующая функция $\tau_\Gamma(n)$ по порядку меньше *) чем n^2. В то же время НС-грамматика, у которой τ имеет порядок (скорости роста) n^2, порождает такой язык; именно такой порядок имеет сигнализирующая функция грамматики, приведенной на стр. 91.

*) Слова «по порядку меньше» приблизительно означают «много меньше», «растет существенно медленнее»; точное определение таково: функция $f(n)$ по порядку меньше функции $g(n)$, если $\lim \frac{f(n)}{g(n)} = 0$ при $n \to \infty$.

Грамматика, у которой временна́я сигнализирующая функция не больше линейной, называется **грамматикой с ограниченным замедлением**. В частности, всякая КС-грамматика является грамматикой с ограниченным замедлением — это следует из неравенства, доказанного на стр. 117—118. Но уже НС-грамматика может не быть грамматикой с ограниченным замедлением.

Аналогичным образом может быть введена и **емкостная сигнализирующая функция** $\sigma_\Gamma(n)$, характеризующая необходимый для вывода объем памяти. Именно,

$$\sigma_\Gamma(n) = \max_{l(x) \leqslant n} \sigma_{\Gamma, x},$$

где Γ, x и n имеют тот же смысл, что и в определении временно́й сигнализирующей функции (стр. 116), а $\sigma_{\Gamma, x}$ есть емкость («объем памяти») наименее емкого вывода цепочки x в Γ. (**Емкостью вывода** мы называем длину самой длинной цепочки, входящей в этот вывод.)

В математической теории грамматик существуют и другие направления исследования, на которых, однако, мы здесь останавливаться не будем.

§ 6. Некоторые другие понятия и проблемы математической лингвистики

Закончив более или менее связное рассмотрение одного из разделов математической лингвистики — теории порождающих грамматик, мы попытаемся теперь дать краткий очерк ряда других разделов, с тем чтобы читатель мог получить более полное представление о всей дисциплине в целом. Этот очерк отнюдь не призван служить обзором хотя бы основных достижений математической лингвистики и никоим образом не претендует на полноту и последовательность. Мы просто коснемся — с очень разной степенью подробности — нескольких отдельных достаточно важных вопросов, не заботясь специально о единстве изложения.

За последние годы в лингвистике сложилось следующее представление об основных направлениях исследования языка и о взаимоотношениях между ними *). Прежде всего различаются:

А) «собственно лингвистические» исследования и, так сказать,

Б) «лингвистические исследования второго порядка». Содержанием первых является описание языка как такового, в частности моделирование речевой деятельности; вторые же направлены на изучение методов исследования языка, в частности на моделирование деятельности лингвистов. Различие между А и Б не совпадает с различием между описательной и теоретической лингвистикой: в А включаются не только описания конкретных языков, но и теория описания языка вообще, построение

*) Это представление мы излагаем в схематизированном и упрощенном виде, чтобы сделать его более отчетливым.

схем конкретных описаний и т. п. Во втором направлении (Б) можно выделить (I) проблемы, связанные с уточнением самих лингвистических понятий,—изучение оснований лингвистики, и (II) проблемы, связанные с уточнением процедур исследования. Эти последние в свою очередь подразделяются на 1) процедуры работы исключительно с текстами — чисто «дешифровочный» подход, и 2) процедуры работы с привлечением информанта — «экспериментальный» подход. (Изложенная классификация лингвистических направлений близка к классификации, предложенной в книге Апресян 1966.)

Намеченную здесь картину можно для наглядности представить так:

Лингвистика.

А. Собственно лингвистика.

Б. Моделирование лингвистических исследований.

I. Основания лингвистики.

II. Процедуры исследования.

(Еще раз напоминаем, что все эти деления весьма приблизительны; четких границ между разделами нет!)

Оказывается, что и в математической лингвистике естественно выделяются аналогичные направления. Поэтому мы будем вести наш обзор, ориентируясь на те же самые рубрики.

Математическая лингвистика также может быть подразделена на:

А) моделирование языков и

Б) моделирование исследования языков (ср. «устройства А и Б» в концепции Н. Хомского — Хомский 1965а, стр. 480—481).

А. Моделирование языков

В данный раздел входит прежде всего теория формальных грамматик. Часть этой теории, связанная с порождающими грамматиками, была изложена в §§ 2—5. Другая часть этой теории относится к распознающим грамматикам (стр. 24). Мы охарактеризуем здесь два класса распознающих грамматик.

категориальные грамматики и автоматы с магазинной памятью.

К этому же разделу можно, по-видимому, отнести те работы, которые обычно объединяются под названием «логический анализ языка» (см. ниже, стр. 153).

Категориальные грамматики (К-грамматики)

Допустим, что перед нами стоит цель построить механическую процедуру синтаксического анализа предложений. Обычное представление о том, как это следует делать, приблизительно таково. Прежде всего надо разбить все словоформы на классы и составить словарь, где каждой словоформе будет приписан ее синтаксический класс (например, *длинная* — $A_{\text{ж,ед,им}}$, *крича* — $V_{\text{деепр}}$ и т. п.). Затем формулируются правила комбинирования синтаксических классов, а именно, указывается, какие классы могут сочетаться с какими и каковы будут классы получающихся словосочетаний. Эти правила можно представить себе, скажем, в виде хорошо известных лингвистам правил НС-анализа, т. е. что-то вроде

$$A + S = S$$

$$S + S_{\text{род}} = S$$

$$V^{tr} + S_{\text{вин}} = V^{itr} \text{ и т. п.}$$

(Заметим, что эти правила по существу суть «обращенные» КС-правила; подробнее об этом см. также ниже, стр. 151.)

Однако представляется заманчивым обойтись без списка правил указанного типа. Для этого можно поступить так: разработать такую систему кодирования синтаксических классов (словоформ и словосочетаний), чтобы возможность комбинирования двух классов, а также класс результирующего сочетания можно было усмотреть непосредственно из кодов исходных классов. Другими словами, все

коды должны иметь определенную «внутреннюю форму», быть «говорящими», и притом, что особенно важно, коды разных классов должны быть согласованы между собой и образовывать алгебраическую систему так, чтобы с помощью несложных операций можно было по кодам комбинирующихся классов механически получать код класса полученной комбинации. Таким образом, здесь вся информация о синтаксической сочетаемости сосредоточена не в списке правил (как это обычно делается в лингвистике и как это было во всех рассмотренных выше грамматиках), а в самих синтаксических кодах слов.

На основе идеи перенесения всей информации о сочетаемости в синтаксический код слова можно построить, вообще говоря, разные классы грамматик. Однако к настоящему времени фактически имеется только один такой класс, а именно, К-грамматики, о которых здесь идет речь.

Сущность К-грамматик состоит в том, что для реализации идеи «говорящего» синтаксического кода привлекается следующее фундаментальное соображение. Все синтаксические классы можно разделить на два типа: некоторые классы считаются основными, или элементарными, а другие рассматриваются как (одноместные) операторы, каждый из которых, будучи применен к какому-либо классу (любого типа), дает снова синтаксический класс*). При этом применяется особый способ обозначения операторов, который можно проиллюстрировать таким примером: пусть имеются элементарные синтаксические классы $S_{им}$ (существительное в им. пад.) и ПРЕДЛ (предложение); тогда синтаксический класс, присоединяемый к $S_{им}$ справа и дающий предложение (т. е. непереходный глагол типа *спит*, *ходит*, *плачет*,...: *Море спит* и т. п.), есть оператор, действующий на $S_{им}$ справа и переводящий $S_{им}$ в ПРЕДЛ; такой оператор обозначается

*) Результирующий класс может в частном случае совпадать с исходным.

$[S_{им} \setminus$ ПРЕДЛ$]$. Действие этого оператора на $S_{им}$ можно трактовать как сокращение дроби: $S_{им}[S_{им} \setminus$ПРЕДЛ$]$ = ПРЕДЛ. Класс глаголов типа *существует*, *имеется*, *появляется*, ... должен рассматриваться как оператор, действующий на $S_{им}$ слева и дающий ПРЕДЛ (*Существует закон* и т. п.); такой оператор обозначается [ПРЕДЛ $/ S_{им}]$ и тоже «сокращается» с $S_{им}$: [ПРЕДЛ $/ S_{им}]S_{им}$= ПРЕДЛ. Прилагательное можно обозначить как $[S/S]$ (препозитивное) или $[S\setminus S]$ (постпозитивное); соответствующие сокращения имеют вид: $[S / S]S = S$ и $S[S \setminus S] = S$.

Уже из этих упрощенных примеров видно, как должен выполняться анализ предложения: всем словоформам приписываются с помощью словаря коды указанного типа (в случае синтаксической неоднозначности словоформы ей приписывается несколько кодов); затем осуществляются все возможные сокращения, и если в результате получится символ предложения ПРЕДЛ, то это означает, что анализируемое предложение грамматически правильно, а тот способ сокращения, который привел к символу ПРЕДЛ, задает синтаксический анализ этого предложения.

Понятие К-грамматики было введено И. Бар-Хиллелом (Bar-Hillel — Gaifman — Shamir 1960), существенно опиравшимся на работы К. Айдукевича. Содержательное изложение соответствующих идей дано в Bar-Hillel 1953; см. также Бар-Хиллел 1964 и Гладкий 1966.

Теперь мы можем перейти к определению К-грамматики. Для этого надо ввести следующие понятия.

1. Рассматривается конечное множество символов V — «основной словарь»; его роль и интерпретация таковы же, как и для порождающих грамматик.

2. Далее, имеется конечный набор W э л е м е н т а р н ы х (синтаксических) к а т е г о р и й — ЭК; формально ЭК — это просто символ из W, а содержательно — это класс слов (точнее, словоформ или

морфем) или словосочетаний. Из ЭК строятся категории по следующим правилам:

1) Всякая ЭК есть категория;

2) если Φ и Ψ — категории, то [Φ\Ψ] и [Φ\Ψ] — тоже категории;

3) никаких других категорий нет (об интерпретации выражений [Φ/Ψ] и [Φ\Ψ] см. выше).

3. Имеется также функция *f*, которая каждому основному символу ставит в соответствие одну или несколько (конечное число!) категорий. Содержательно эту приписывающую функцию можно представлять себе как задание в словаре при словах их синтаксических классов (нескольких — в случае омонимии: *течь* — существительное или глагол).

4. Среди ЭК выделяется главная категория ***F***; это, так сказать, «конечный» символ, роль которого в некотором смысле противоположна роли начального символа ***I*** порождающей грамматики (из ***I*** развертывается порождаемая цепочка, к ***F*** свертывается распознаваемая цепочка).

К-грамматика представляет собой совокупность описанных четырех объектов, т. е., говоря формально, упорядоченную четверку ⟨***V***, ***W***, *f*, ***F***⟩.

Укажем, как работает такая грамматика, введя предварительно еще одно понятие: сокращение цепочек категорий. Под непосредственным сокращением понимается одна из двух операций:

либо некоторое вхождение цепочки [Φ/Ψ] Ψ заменяется на Φ (правое сокращение),

либо некоторое вхождение цепочки [Φ[Φ\Ψ] заменяется на Ψ (левое сокращение).

Запись [Φ/Ψ] читается «Φ над Ψ», а Φ\Ψ]— «Φ под Ψ», и непосредственное сокращение удобно представлять себе как сокращение дробей с тем существенным отличием, что здесь взаимное расположение «сомножителей» небезразлично. Категория [Φ/Ψ] — это категория, приписываемая такой цепочке, которая, находясь слева от цепочки с категорией Ψ, образует вместе с нею цепочку,

имеющую в целом категорию Ф:

$$\ldots \overbrace{|[\Phi/\Psi]| \; |\Psi|}^{\Phi} \ldots$$

Иначе говоря, $[\Phi/\Psi]$, как уже указывалось, есть оператор, который, действуя на Ψ слева, превращает ее в Ф.

Совершенно аналогично интерпретируется категория $[\Phi\setminus\Psi]$, но только здесь идет речь о присоединении, во-первых, не к Ψ, а к Ф, а во-вторых, не слева, а справа.

Определим с о к р а щ е н и е (не обязательно непосредственное): цепочка категорий α сокращается до цепочки категорий β, если β получается из α последовательностью непосредственных сокращений.

Например, цепочка

$$\alpha = [[[X\setminus Y]/X]/Z]Z[Y/Y][[Y/Y]\setminus X][Z/X]X$$

сокращается до цепочки $\beta = [X\setminus Y]Z$ путем четырех непосредственных сокращений:

$$\alpha = [[\underbrace{[X\setminus Y]/X}_{\gamma}]/Z]Z[Y/Y][[Y/Y]\setminus X][Z/X]X$$

1. $[\gamma/Z]Z \to \gamma$

$$\underbrace{[[X\setminus Y]/X]}_{\gamma}\underbrace{[Y/Y]}_{\delta}[\underbrace{[Y/Y]}_{\delta}\setminus X][Z/X]X$$

2. $\delta[\delta\setminus X] \to X$

$$[\underbrace{[X\setminus Y]}_{\varepsilon}/X]X[Z/X]X$$

3. $[\varepsilon/X]X \to \varepsilon$

$$\underbrace{[X\setminus Y]}_{\varepsilon}[Z/X]X$$

4. $[Z/X]X \to Z$

$$\beta = [X\setminus Y]Z$$

Отметим, что одну и ту же цепочку категорий

можно, вообще говоря, сокращать разными способами, применяя непосредственные сокращения в разном порядке. Соответствующий пример будет рассмотрен ниже, на стр. 132.

Теперь мы перейдем к описанию работы К-грамматики, причем для большей наглядности мы сначала построим пример такой грамматики.

Основной словарь V этой грамматики (G_1) состоит из русских словоформ; в качестве элементарных категорий выберем следующие:

ПРЕДЛ — предложение;

$V^{itr}_{x,\,y,\,w,\,v}$, где V^{itr} — непереходный глагол, а x, y, w и v — переменные, обозначающие соответственно: x — род, y — число, w — лицо и v — время; ср. стр. 57. Таким образом, $V^{itr}_{x,\,y,\,w,\,v}$ — не одна ЭК, а сокращенное обозначение для 54 ЭК: $54 = 3$ (рода) $\times$ $\times\ 2$ (числа) $\times\ 3$ (лица) $\times\ 3$ (времени);

$V^{tr}_{x,\,y,\,w,\,v}$, где V^{tr} — переходный глагол, а x, y, w, v имеют тот же смысл, что и выше;

Pr_i, где Pr — общее обозначение для предлога, а i — переменная, обозначающая конкретный предлог: *без*, *в*, *на*, *от*,... Pr_i — также сокращенное обозначение для нескольких десятков ЭК.

Категорию ПРЕДЛ будем считать главной.

Разумеется, мы будем иметь дело не со всеми русскими словоформами, а лишь с несколькими: *лепила*, *на*, *метель*, *кружкй*, *стрелы*, *стекле*, *и*. Только для них мы и определим значения приписывающей функции f, причем укажем не все ее возможные значения*), а только те, которые нужны для нашего примера:

1) f (*лепила*) $= V^{tr}_{\text{ж, ед,3, прош}}$

2) f (*на*) $= Pr_{\text{на}}$

3) f (*метель*) $=$ [ПРЕДЛ$\diagdown V^{itr}_{\text{ж, ед, 3},\,v}$]

---

*) Напомним, что f — многозначная функция (стр. 125). Так, словоформе *метель* она должна была бы сопоставлять много разных категорий, см. ниже.

4, 5) f (*кружки*) $= f$(*стрелы*) $=$

$$= [V^{tr}_{x,\,y,\,w,\,v} \diagdown V^{itr}_{x,\,y,\,w,\,v}],$$

6) f (*стекле*) $= [Pr_{\text{на}} \diagdown [V^{tr}_{x,\,y,\,w,\,v} \diagdown V^{tr}_{x,\,y,\,w,\,v}]]$

7) $f\,(и) = [[[V^{tr}_{x,\,y,\,w,\,v} \diagdown V^{itr}_{x,\,y,\,w,\,v}] \diagdown [V^{tr}_{x,\,y,\,w,\,v} \diagdown$

$$\diagdown V^{itr}_{x,\,y,\,w,\,v}]] \diagup [V^{tr}_{x,\,y,\,w,\,v} \diagdown V^{itr}_{x,\,y,\,w,\,v}]].$$

Категория, приписанная словоформе *метель*, есть оператор, который, действуя слева на непереходный глагол соответствующего рода, числа и лица, дает предложение; другими словами, это категория «Подлежащее, стоящее перед сказуемым» («Левое подлежащее»). Чтобы учесть возможность постановки подлежащего после сказуемого, нам пришлось бы приписать словоформе *метель* еще одну категорию: $[V_{\text{ж, ед, 3},\,v} \diagdown \text{ПРЕДЛ}]$ *). А если бы мы захотели учесть также и омонимию падежей — форма *метель* может быть и прямым дополнением, правым или левым, — то мы должны были бы приписать этой словоформе еще две группы категорий:

$$[V^{tr}_{x,\,y,\,w,\,v} \diagdown V^{itr}_{x,\,y,\,w,\,v}] \text{ и } [V^{itr}_{x,\,y,\,w,\,v} \diagup V^{tr}_{x,\,y,\,w,\,v}].$$

Однако и это еще не все: чтобы отразить способность формы *метель* сочетаться с разными предлогами (*в метель*, *за метель*...), ей надо было бы приписать еще целый ряд категорий.

Сказанное здесь относится и к случаям 4) —7): этим словоформам приписаны не все категории, которые они имеют в русском языке, а только те, которые нужны для нашего примера. А именно, словоформам *кружки* и *стрелы* приписана категория «Правое прямое дополнение» **), словоформе *стекле* — категория «Дополнение предлога *на*» (оператор, действующий на *на* справа и дающий определение к переходному ***) глаголу), а словофор-

*) Строго говоря, это не одна категория, а группа из трех категорий: переменная v может принимать три значения.

**) «Правое прямое дополнение» есть оператор, действующий на переходный глагол справа и превращающий его в глагольную группу, эквивалентную непереходному глаголу.

***) Определение к непереходному глаголу должно иметь другую категорию.

ме *и* — категория «Союз, соединяющий правые прямые дополнения» (оператор, который, оказавшись между двумя правыми прямыми дополнениями, дает снова правое прямое дополнение).

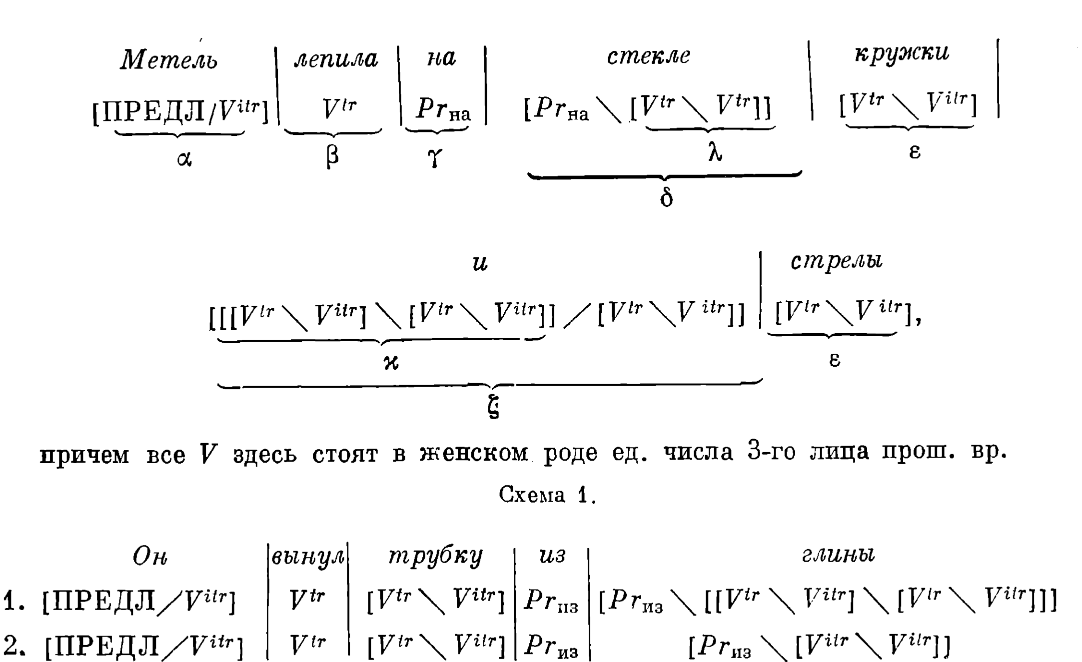
Итак, грамматика G_1 построена. На ее примере мы можем описать принцип работы К-грамматик вообще.

Пусть имеется цепочка $x = a_1 a_2 a_3 \ldots a_k$ из символов основного словаря некоторой К-грамматики G_1. Приписывающая функция f этой грамматики позволяет сопоставить цепочке x цепочку категорий $\xi = f(a_1)f(a_2)f(a_3)\ldots f(a_k)$ (в общем случае таких цепочек может быть несколько — за счет многозначности функции f, т. е. омонимичности некоторых a_i). Если эта цепочка (в общем случае — хотя бы одна из таких цепочек) сокращается до одной категории Ф, то мы будем говорить, что грамматика G_1 приписывает цепочке x категорию Ф.

Рассмотрим, например, следующую цепочку русских словоформ: *Метель лепила на стекле кружки и стрелы*. В грамматике G_1 этой цепочке может быть сопоставлена цепочка категорий, изображенная на схеме 1 (см. стр. 130). Данную цепочку категорий можно сократить следующим образом:

1) $\zeta\varepsilon \to \varkappa$, 2) $\varepsilon\varkappa \to \varepsilon$, 3) $\gamma\delta \to \lambda$, 4) $\beta\lambda \to V^{tr}_{\text{ж,ед,3,прош}}$, 5) $V^{tr}_{\text{ж,ед,3,прош}} \quad \varepsilon \to V^{itr}_{\text{ж,ед,3,прош}}$, 6) $\alpha V^{itr}_{\text{ж,ед,3,прош}} \to$ $\to$ ПРЕДЛ. Это значит, что G_1 приписывает цепочке *Метель лепила на стекле кружки и стрелы* категорию ПРЕДЛ, т. е. «признает» эту цепочку предложением. Соответственно цепочке *кружки и стрелы* приписывается категория $\varepsilon = [V^{tr}_{\text{ж,ед,3,прош}} \diagdown V^{itr}_{\text{ж,ед,3,прош}}]$ («Правое прямое дополнение»), цепочке *лепила на стекле* — категория $V^{tr}_{\text{ж,ед,3,прош}}$, а цепочке *лепила на стекле кружки и стрелы* — категория $V^{itr}_{\text{ж,ед,3,прош}}$. В то же время нетрудно видеть, что таким цепочкам, как *метель лепила на*, *стекле кружки и* или *на и*, *на лепила*, *лепила стекле* и т. п., данная грамматика не приписывает никаких категорий.

Таким образом, для произвольной цепочки (из символов словаря V) К-грамматика позволяет узнать,



причем все V здесь стоят в женском роде ед. числа 3-го лица прош. вр.

Схема 1.

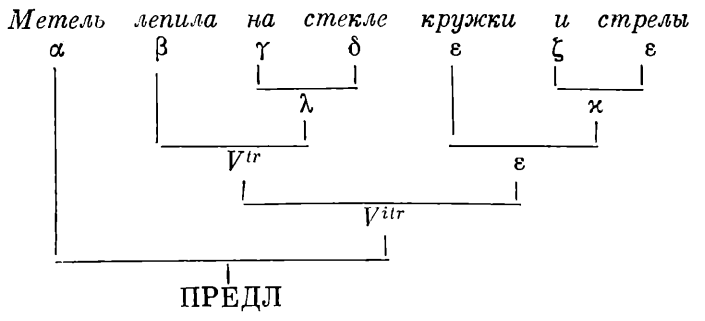
	Он	*вынул*	*трубку*	*из*	*глины*
1.	[ПРЕДЛ/V^{itr}]	V^{tr}	[$V^{tr} \setminus V^{itr}$]	$Pr_{\text{из}}$	[$Pr_{\text{из}} \setminus [[V^{tr} \setminus V^{itr}] \setminus [V^{tr} \setminus V^{itr}]]$]
2.	[ПРЕДЛ/V^{itr}]	V^{tr}	[$V^{tr} \setminus V^{itr}$]	$Pr_{\text{из}}$	[$Pr_{\text{из}} \setminus [V^{itr} \setminus V^{itr}]$]

Схема 2.

какими категориями характеризуется эта цепочка и характеризуется ли она вообще какими бы то ни было категориями. В частности, для каждой цепочки К-грамматика позволяет узнать, является ли эта цепочка правильным предложением, т. е. позволяет распознавать грамматическую правильность. Более того, если цепочка оказывается грамматически правильным предложением, то К-грамматика выделяет в нем словосочетания, т. е. составляющие (в обычном смысле этого слова). Способ построения дерева составляющих по записи процесса сокращения исходной цепочки категорий крайне прост: составляющей будет каждая подцепочка, для которой соответствующая цепочка категорий на некотором шаге процесса сокращения сокращается до одной категории; эта категория может рассматриваться как синтаксический тип данной составляющей. В нашем примере имеется шесть составляющих (не считая одноэлементных), соответствующих шести шагам процесса сокращения:

1) *и стрелы* (типа $\varkappa$),
2) *кружки и стрелы* (типа ε),
3) *на стекле* (типа λ),
4) *лепила на стекле* (типа $V^{tr}_{\text{ж, ед, 3, прош}}$),
5) *лепила на стекле кружки и стрелы* (типа $V^{itr}_{\text{ж, ед, 3, прош}}$),
6) вся фраза целиком (типа ПРЕДЛ).

Эта система составляющих представляется привычным деревом НС:



В случае синтаксически неоднозначного предложения К-грамматика может сопоставлять ему разные системы составляющих, т. е. давать разные анализы. Во-первых, это бывает потому, что соответствующая предложению цепочка категорий может сокращаться разными способами, например:

... *медленные* *протоны* *и* *нейтроны*

$$\underbrace{[S/S]}_{\alpha}\qquad S\qquad \underbrace{[\underbrace{[S\diagdown S]}_{\beta}/S]}_{\gamma}\qquad S$$

Первый способ сокращения: $\alpha S\rightarrow S$, $\gamma S\rightarrow\beta$, $S\beta\rightarrow S$, что дает систему составляющих

... (*медленные протоны*) *и нейтроны*.

Второй способ сокращения: $\gamma S\rightarrow\beta$, $S\beta\rightarrow S$, $\alpha S\rightarrow$ $\rightarrow S$, что дает... *медленные* (*протоны и нейтроны*).

Во-вторых, предложению может быть приписано несколько цепочек категорий и это также может привести к разным системам составляющих, см. схему 2 на стр. 130.



Для первой цепочки категорий получается разложение (*он* (*вынул* (*трубку* (*из глины*)))), т. е. трубка была глиняная, а для второй — (*он* ((*вынул трубку*) (*из глины*))), т. е. трубка была вынута из глины. Все необходимые выкладки читателю рекомендуется выполнить самостоятельно — в качестве упражнения *).

С каждой К-грамматикой естественно связывается множество тех цепочек, которые эта грамматика признает предложениями, т. е. приписывает им категорию ПРЕДЛ. Это множество называется *языком, определяемым К-грамматикой*, или *К-языком*.

Возникает вопрос: каково соотношение между языками, определяемыми К-грамматиками, и язы-

*) Здесь фактически затронута целая новая область синтаксической теории, а именно, синтаксическая неоднозначность: ее природа, ее различные типы, ее соотношение с семантической неоднозначностью и т. п. Однако в рамках настоящей книги мы не имеем возможности останавливаться на этих вопросах.

ками, порождаемыми порождающими грамматиками? Легко видеть, что всякий К-язык есть КС-язык и, более того, что для всякой К-грамматики можно

$T_{2.1}$ построить эквивалентную ей КС-грамматику (т. е. порождающую в точности тот же язык, который определяется исходной К-грамматикой). Наметим здесь идею очень простого доказательства этого факта: если $G=\langle V, W, f, \text{ПРЕДЛ}\rangle$ есть К-грамматика, то соответствующая КС-грамматика Γ строится так: 1) основным словарем в Γ будет V; 2) вспомогательным словарем в Γ будет множество всех категорий, являющихся значениями приписывающей функции f или их частями *); 3) начальным символом служит ПРЕДЛ; 4) правила грамматики Γ будут двух типов: во-первых, правила вида $\Psi\to\Phi[\Phi\diagdown\Psi]$ и $\Phi\to[\Phi/\Psi]\Psi$, где $[\Phi\diagdown\Psi]$ и $[\Phi/\Psi]$ — произвольные составные категории из вспомогательного словаря грамматики Γ (таких категорий заведомо конечное число — см. стр. 125); во-вторых, правила вида $f(a)\to a$, где a — основной (терминальный) символ, а $f(a)$ — произвольное значение приписывающей функции от a. Эквивалентность построенной так КС-грамматики Γ и исходной К-грамматики G почти очевидна; мы позволим себе опустить формальное доказательство.

Оказывается, что верно и обратное: всякий КС-язык есть К-язык, причем для всякой КС-грам-

$T_{2.2}$ матики можно построить эквивалентную ей К-грамматику.

ЛВ Из сказанного непосредственно следует, что класс К-языков в точности совпадает с классом КС-языков. Поэтому вопрос о принципиальной пригодности К-грамматик для описания естественных языков решается так же, как для КС-грамматик (см. стр. 93 и сл.). Что же касается практического удобства, то в двух очень существенных отношениях К-грамматики явно неудобны.

---

*) Частью составной категории является любая категория, входящая в ее состав, например, частями категории $[[X\diagdown[Y/X]]/[Y/Z]]$ являются (не считая ее самой) следующие категории: $[X\diagdown[Y/X]]$, $[Y/X]$, $[Y/Z]$, X, Y, Z.

1) Применение их к естественным языкам, в особенности же к языкам с развитой морфологией (например, к русскому), требует введения огромного количества составных категорий, и притом необычайно громоздких — ср. категории для *и* в примере на стр. 130. Что еще хуже, эти категории используются весьма неэкономно: почти каждому слову, даже если оно не омонимично (т. е. не типа *печь* или англ. [*to*] *work* — [*the*] *work*), сопоставляется очень много разных категорий. Мы уже говорили об этом применительно к форме *метель* (стр. 128); обратим внимание еще на союз *и*: на стр. 128 ему уже приписано 54 категории, однако эта группа категорий годится только для случая, когда *и* соединяет правые прямые дополнения! Для всех прочих многочисленных случаев (левые прямые дополнения, правые предложные группы с предлогом *в*, то же с предлогами *без*, *на* и т. п., переходные глаголы, левые определения к глаголу,..., целые предложения) союзу *и* надо приписать особые группы категорий, вообще говоря, такого же объема. В результате союзу *и* будет сопоставлено несколько тысяч категорий.

Очевидно, что если большинству слов сопоставляется (каждому) так много категорий, то число цепочек категорий для одной цепочки слов астрономически велико.

Обратим внимание на то, что в нашем примере нет прилагательных. Это не случайно: если бы при выбранном нами принципе введения элементарных категорий пришлось бы сопоставлять категории и прилагательным, это повело бы к еще большему увеличению числа категорий на слово. Чтобы разъяснить это, нужно слишком много места; читатель может попытаться сам приписать прилагательным нужные категории в рамках предлагаемой грамматики (это потребует некоторых изменений в уже введенных категориях; основная трудность — обеспечить согласование прилагательных с существительными в роде). Сходные трудности, связанные с приписыванием категорий английскому герундию

—например, в *playing cards is fun*—отметил И. Бар-Хиллел (Bar-Hillel 1960).

2) В К-грамматиках синтаксические категории слишком тесно связаны с порядком слов: левое и правое подлежащее, левое и правое дополнение — совершенно различные категории, что резко противоречит интуиции.

В то же время К-грамматики имеют три безусловных достоинства:

1) К-грамматики фактически не имеют правил (за исключением общих для всех грамматик правил сокращения, которых всего два). Все необходимые сведения о синтаксических свойствах слов содержатся здесь только в «словарных статьях», т. е. даются исключительно приписывающей функцией, ср. стр. 125.

2) К-грамматика предполагает весьма тонкий анализ синтаксических свойств отдельных словоформ, т. е. детальное различение их синтаксических функций. При этом для каждой «микрофункции» имеется отдельное эксплицитное выражение — своя категория.

Обратной стороной обоих достоинств как раз и является громоздкость и многочисленность категорий.

3) К-грамматика позволяет получать для фразы не только систему (и дерево) составляющих, но и систему (дерево) зависимостей *). (Подчеркнем, что КС-грамматика, вообще говоря, не дает единственного дерева зависимостей даже для синтаксически однозначных предложений.) Это свойство делает К-грамматики интересными с точки зрения автоматического анализа и перевода текстов, где удобно иметь возможность одновременного представления структуры текста и в терминах НС, и в терминах зависимостей **).

---

*) Относительно обоих методов представления синтаксической структуры см. Падучева 1964б; Гладкий 1966, стр. 8—29; Hays 1961.

**) Ряд алгоритмов анализа текстов, основанных на тех или иных модификациях К-грамматик, упоминается в Bobrow 1963.

В то же время для подобных практических целей К-грамматики в чистом виде — в связи с указанными выше недостатками — по-видимому, мало пригодны.

Автоматы с магазинной памятью (М-автоматы)

Прежде всего мы попытаемся дать содержательное представление об автоматах с магазинной памятью с помощью примера. Использование магазинной памяти положено в основу одного класса алгоритмов автоматического синтаксического анализа*) — так называемых предсказуемостных, или предиктивных, анализаторов (П-анализаторов; Kuno 1963, Kuno — Oettinger 1963, Plath 1963).

Ниже описывается П-анализатор не совсем такого вида, как в указанных работах. Поскольку в отличие от авторов этих работ мы не связаны требованиями практического удобства обработки реальных текстов, но должны заботиться о максимальной простоте и общности изложения, оказалось целесообразным придать П-анализатору несколько иную форму.



Главное отличие предлагаемого ниже варианта П-анализатора от «традиционных» вариантов состоит в том, что обычно П-анализаторы трактуют как алгоритмы синтаксического анализа, мы же рассматриваем их здесь как исчисления (грамматики). Иначе говоря, в обычном П-анализаторе каждый шаг однозначно определяется предыдущим ходом процесса, тогда как в нашем варианте для каждого шага имеется, вообще говоря, много возможностей («разрешенных вариантов»), из кото-

*) Алгоритм автоматического синтаксического анализа — это система (правил), способная приписывать фразам данного языка их синтаксические структуры. Такие алгоритмы являются важнейшими частями систем автоматического перевода. Подробнее об автоматическом синтаксическом анализе см. Мельчук 1963; 1964, Tosh 1965.

рых можно выбирать любую (ср. замечание о соотношении грамматики и алгоритма на стр. 45—46).

Общая идея предсказуемостного анализа заключается, грубо говоря, в следующем. Фраза обрабатывается слово за словом в одном направлении — слева направо. Для каждого очередного слова формулируется «синтаксическое предсказание» (СП) — предсказывается, какая конструкция (составляющая) может следовать в данной фразе за этим словом. Если следующее слово удовлетворяет этому СП, анализ продолжается, в противном случае прекращается. Поскольку слово может предсказывать много разных составляющих, образуется много «дорожек» анализа, причем в случае синтаксически однозначной фразы анализ доводится до конца только по одной дорожке.

В результате анализа фразе сопоставляется система составляющих («анализ по НС»); границы составляющей будут изображаться с помощью скобок, помечаемых символом этой составляющей.

Процесс работы П-анализатора мы опишем на примере анализа фразы: *Маленький мальчик на далеком от города полустанке ждал приезда родителей*. Для обработки этой фразы необходимо располагать рядом сведений о русском языке; эти сведения размещаются в следующих двух компонентах П-анализатора:

1) Словарь (или словарь + алгоритм морфологического анализа), который сопоставляет каждой словоформе ее синтаксический код: часть речи, падеж, род, лицо, одушевленность и т. п., например, для *маленький* — $A_{\text{м,ед,им.-вин}}$, для *на* — $Pr_{\text{вин/предл}}$ и т. д. Этот компонент — аналог приписывающей функции категориальной грамматики, см. стр. 125 — характерен не только для П-анализатора: подобный словарь в том или ином виде необходим в любых распознающих грамматиках (а фактически — и в порождающих, ср. пример на стр. 57, где роль словаря играют правила группы IV).

2) Специфической особенностью П-анализатора являются две синтаксические таблицы

А и Ω (см. стр. 139—140), содержащие все необходимые сведения о синтаксисе данного языка.

Таблица А, описывающая возможные «начала» разных составляющих, имеет три столбца.

В столбце I перечисляются все типы составляющих *), принятые в данном описании данного языка, например «именная группа», «группа сказуемого», «предложная группа» и т. д.

В столбце II для каждого типа составляющих C перечисляются все классы словоформ w_1, w_2, w_3,… …, w_n, которыми может начинаться составляющая данного типа, например, для именной группы в им. падеже — 1) наречие (*очень хорошая книга*), 2) $A_{\text{им}}$ (*маленький мальчик*), 3) $S_{\text{им}}$ (*таблица*) и т. д.

И наконец, в столбце III содержатся сами синтаксические предсказания (СП). А именно, для каждой пары $\langle C, w_i \rangle$ указывается, какие составляющие (одна или несколько) могут следовать за w_i внутри составляющей C. Например, одним из синтаксических предсказаний для словоформы типа *приезд*, *прыжок* и т. п., начинающей «именную группу», будет «именная группа в род. пад», что содержательно означает следующее: такая словоформа способна иметь дополнение в род. падеже в отличие, например, от *он* или *нас*, при которых именной группы в род. падеже вообще быть не может. Среди СП для прилагательного, начинающего именную группу, есть «предложная группа + именная группа» (см. табл. А, строка 3): это означает, что именная группа может иметь строение «прилагательное + + предложная группа + именная группа», например, *трудная + для нас + алгебраическая задача*.

Для каждой пары $\langle C, w_i \rangle$ имеется, вообще говоря, несколько альтернативных синтаксических предсказаний: например, после словоформы *приезда* в составе именной группы может непосредственно следовать либо предложная группа (*приезда в Москву*),

*) За исключением составляющих, равных отдельной словоформе («атомарных» составляющих) или целому предложению.

Синтаксическая таблица А

Номер строки	I Составляющая	II Класс начальной словоформы	III Синтаксические предсказания	Примеры
1.	ИГ$_{x,y,z}$	*Adv*	ГО$_{x,y,z}$ + + ИГ$_{x,y,z}$	*Очень маленький мальчик...*
2.	ИГ$_{x,y,z}$	$A_{x,y,z}$	ИГ$_{x,y,z}$	*Маленький мальчик . . .*
3.	ИГ$_{x,y,z}$	$A_{x,y,z}$	ПГ + ИГ$_{x,y,z}$	*Бледное от усталости лицо . . .*
4.	ИГ$_{x,y,z}$	$S_{x,y,z}$	ПГ	*Дом с мезонином . . .*
5.	ИГ$_{x,y,z}$	$S_{x,y,z}$	ИГ$_{x',y'}$, род	*Сын наших старых соседей*
6.	ГО$_{x,y,z}$	$A_{x,y,z}$	ПГ	*Близкий к окончанию . . .*
7.	ПГ	Pr_z	ИГ$_{x,y,z}$	*На зеленом лугу . . .*
8.	ГГ$_{x,y,w}$	*Adv*	ГГ$_{x,y,w}$	*Давно мечтает . . .*
9.	ГГ$_{x,y,w}$	Pr_z	ИГ$_{x',y',z}$ + +ГГ$_{x,y,w}$	*На голой вершине стоит*
10.	ГГ$_{x,y,w}$	$V_{x,y}$	ПГ	*. . . стоит за углом*
11.	ГГ$_{x,y,w}$	$V_{x,y,z'}$	ИГ$_{z'}$	*. . . ждет билетов*
12.	ГГ$_{x,y,w}$	$V_{x,y,z',z''}$	ИГ$_{z'}$ + ИГ$_{z''}$	*. . . дает билет подруге*

Пояснения к обозначениям (в таблицах А и Ω):

ИГ — именная группа,
ПГ — предложная группа,
ГГ — глагольная группа,
ГО — группа определения,

Pr_z — предлог, требующий падежа z,
$V_{z'}$ — глагол, требующий дополнения в падеже z',
$V_{z'z''}$ — глагол, требующий двух дополнений: в падеже z' и в падеже z''.

Остальные обозначения имеют тот же смысл, что и выше, стр. 57.

либо именная группа в родительном падеже (*приезда алжирских студентов*), либо придаточное с *который*, а также может ничего не следовать.

Синтаксическая таблица А (как и таблица Ω), носит сугубо иллюстративный характер и вовсе не претендует на сколько-нибудь адекватное представление русского синтаксиса; достаточно полная таблица А должна быть во много раз больше (синтаксическая таблица для русского языка, приведенная в Plath 1963, имеет 2344 строки и гораздо более сложное строение).

Следует иметь в виду, что, как и раньше (ср. стр. 55), здесь использованы сокращенные обозначения с помощью переменных индексов. Таким образом, каждая строка таблицы А (и таблицы Ω) фактически является записью многих строк.

Таблица Ω описывает возможные «концы» разных составляющих и имеет всего два столбца.

В столбце I, как и в таблице А, перечисляются все типы составляющих, кроме атомарных.

В столбце II для каждого типа составляющих С перечисляются все классы словоформ (в нашем примере приводятся не все!), которыми составляющая данного типа может оканчиваться, например, для «именной группы» — S и т. п.

В нашем примере таблица Ω имеет следующий вид:

Синтаксическая таблица Ω

I Составляющая	II Класс замыкающей словоформы	Примеры
1. $\text{ИГ}_{x,y,z}$	$S_{x,'y,'z'}$	. . . *берег с небольшим лесочком* .
2. $\text{ГО}_{x,y,z}$	$S_{x,'y,'z'}$	. . . *важные для читателя* .
3. $\text{ГО}_{x,y,z}$	$A_{x,y,z}$	. . . *давно забытый* .
4. ПГ	$S_{x,y,z}$	. . . *на зеленом лугу* .
5. $\text{ГГ}_{x,y}$	$V_{x,y}$	. . . *усердно работает* .
6. $\text{ГГ}_{x,y}$	$S_{x,'y,'z'}$	. . . *работает над книгой*.
7. ПРЕДЛ	$S_{x,y,z}$	*Лодка колотится в сонной груди* .
8. ПРЕДЛ	$V_{x,y}$	*Сердце с домом разлучается*.

Кроме обоих описанных компонентов, содержащих все необходимые сведения о синтаксисе данного языка, П-анализатор имеет еще третий, наиболее существенный компонент — особым образом организованную «рабочую зону», называемую накопителем синтаксических предсказаний (НСП). В отличие от словаря и обеих синтаксических таблиц, которые используются как пассивные справочники (их содержимое в процессе анализа не меняется), НСП есть именно то место, где происходит переработка всех промежуточных результатов. НСП состоит из неограниченного числа *магазинов**), которые можно представлять себе как устройства, аналогичные винтовочному магазину или автомату, подающему бутерброды: в магазин можно вводить (записывать) только по одному символу и только с одного конца, скажем, сверху, и только с того же конца можно «вынимать» символы из магазина (стирать их), так что символ, введенный в магазин последним, вынимается из него первым (last-in-first-out principle); при этом магазин устроен так («снабжен подающей пружиной»), что, когда в него вводится новый символ, все символы, бывшие там ранее, опускаются вниз («утапливаются»), а когда верхний символ стирается (вынимается), остальные символы автоматически подымаются на шаг вверх (рис. на стр. 142).

Теперь мы можем перейти к изложению самого процесса анализа фразы с помощью П-анализатора.

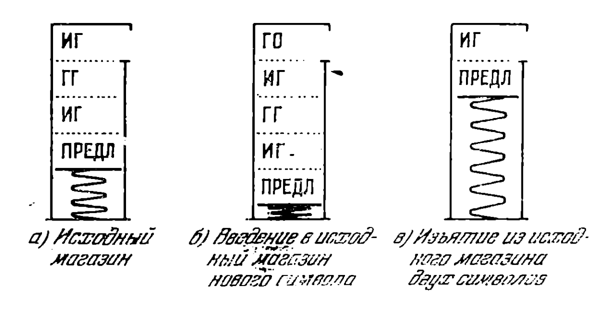
Прежде всего следует подчеркнуть, что как в силу грамматической неоднозначности многих

*) Это означает, что данный П-анализатор имеет потенциально бесконечное множество магазинов, т. е. сколько бы магазинов ни понадобилось в процессе анализа, столько их и можно иметь в распоряжении. Нужное количество магазинов зависит от длины и синтаксических особенностей фразы: каждый магазин соответствует одному варианту анализа, причем отдельные магазины в принципе нужны не только для правильных (доходящих до конца фразы) вариантов, но и для неправильных (обрывающихся где-то по дороге).

Сами магазины также считаются имеющими неограниченную вместимость, что связано с фундаментальным допущением о неограниченности длин фраз естественного языка (см. стр. 59—60).

словоформ (*печь* — $V_{inf}/S_{\text{им-вин}}$, *линии* — $S_{\text{мн, им-вин}}/$ $/S_{\text{ед,род}}/S_{\text{ед,предл}}$ и т. д.), так и в силу наличия у словоформы в большинстве случаев нескольких СП каждый шаг анализа допускает, вообще говоря, более одного продолжения. Это приводит ко многим вариантам, или «дорожкам», анализа.

Однако в некоторых вариантах возможна ситуация, когда на некотором шаге не оказывается ни одного допустимого продолжения, а обработка



а) Исходный магазин б) Введение в исходный магазин нового символа в) Изъятие из исходного магазина двух символов

фразы еще не закончена. Это означает, что либо анализируемая фраза синтаксически неправильна, либо (если она правильна) на одном из предыдущих шагов было выбрано «нежелательное» продолжение, что завело анализ в тупик. Таким образом, правильность дорожки (т. е. варианта анализа) сигнализируется достижением конца фразы, неправильность — невозможностью дойти до конца фразы.

Каждой дорожке отводится один магазин. Этот магазин играет роль памяти (отсюда и название «автоматы с магазинной памятью»), в которой хранятся синтаксические предсказания, соответствующие данной дорожке.

При рассмотрении примера мы проследим только одну дорожку анализа, соответствующую правильному синтаксическому толкованию приведенной на стр. 137 фразы (эта фраза имеет и другое правильное толкование, см. ниже стр. 147).

До начала анализа перед первым подлежащим обработке словом открывается скобка с пометкой

«ПРЕДЛ», а в магазины помещаются предсказания $ИГ_{м, ед, им}$ + $ГГ_{м, ед, 3}$ + ПРЕДЛ (табл. 0 на стр. 146), поскольку a priori можно предполагать, что анализируемый отрезок представляет собой предложение (ПРЕДЛ) и состоит из группы подлежащего ($ИГ_{им}$) и группы сказуемого (ГГ), причем $ИГ_{им}$ стоит левее, чем ГГ. Это предположение, разумеется, не единственное; в другие магазины могут быть помещены иные априорные наборы предсказаний — хотя бы на подлежащее (и, соответственно, сказуемое) женского или среднего рода или множественного числа, — которые, однако, для нашей фразы не дадут правильных анализов.

Каждый шаг анализа состоит в обработке одного слова. Эта обработка складывается из следующих операций (напоминаем — ср. стр. 136, — что здесь дается не алгоритм анализа, а набор р а з р е ш е н н ы х операций, так что на каждом шаге можно выполнять любое из возможных в этом случае действий):

1. Сравнить синтаксический код слова с рабочим (верхним) предсказанием в магазине, т. е. искать соответствующую пару в таблице А и тем самым проверить, может ли данное слово начинать предсказанную составляющую. Если нет, продолжение анализа невозможно.

2. Если да, начало этой составляющей отмечается левой скобкой, и дальше можно делать одно из двух:

2а. В предположении, что начатая составляющая продолжается, выбрать одно (любое) из СП рабочего слова, поместить его в магазин (поверх хранящихся там предсказаний) *) и перейти к следующему шагу, т. е. к очередному слову.

2б. В предположении, что начатая составляющая заканчивается рабочим словом, искать в таблице Ω пару ⟨рабочее СП, синтаксический код рабочего слова⟩ и тем самым проверить, может ли данное

*) Если это СП состоит более чем из одной синтаксической группы (см. столбец III таблицы А, строки 1, 3, 9 и 12), то группы помещаются в магазин справа налево, так что самой верхней оказывается самая левая группа.

слово заканчивать эту составляющую. Если нет, продолжение анализа невозможно. Если да, конец составляющей отмечается правой скобкой и соответствующее (рабочее) СП убирается из магазина. Операция 2б может быть повторена сколько угодно раз (поскольку рабочее слово может заканчивать не только данную составляющую, но и более крупную, которая была начата ранее и включает данную); затем следует переходить к очередному шагу анализа.

Анализ считается законченным, а результат его правильным, если после обработки последнего слова фразы магазин оказывается пустым. Итак, начинаем описание анализа нашей фразы.

1-й шаг. Рассматриваем слово *маленький*; рабочее (верхнее в магазине) СП$=\text{ИГ}_{\text{м,ед,им}}$. Поскольку пара $\langle \text{ИГ}_{\text{м,ед,им}}, \text{А}_{\text{м,ед,им}} \rangle$ есть в табл. А (строки 2, 3), мы заключаем, что словоформа типа *маленький* может начинать предсказанную составляющую. Ставим перед *маленький* левую скобку, помеченную символом «$\text{ИГ}_{\text{м,ед,им}}$», и выбираем путь 2а, т. е. берем в столбце III одно из СП данной пары, а именно $\text{ИГ}_{\text{м,ед,им}}$, и помещаем его в магазин; теперь магазин будет содержать предсказания $\text{ИГ}_{\text{м,ед,им}} + \text{ИГ}_{\text{м,ед,им}} + \text{ГГ}_{\text{м,ед,3}} + \text{ПРЕДЛ}$ (см. табл. 1, стр. 146).

2-й шаг. Рабочее слово — *мальчик*, рабочая пара — $\langle \text{ИГ}_{\text{м,ед,им}}, S_{\text{м,ед,им}} \rangle$. Эта пара содержится в табл. А (строки 4, 5), и мы, поставив перед *мальчик* соответствующую скобку, выбираем на этот раз путь 2б: ищем рабочую пару в таблице Ω, найдя ее там (строка 1), ставим после *мальчик* правую скобку («$\text{ИГ}_{\text{м,ед,им}}$») и убираем рабочее предсказание из магазина. Повторяем операцию 2б еще раз: снова находим в Ω рабочую пару $\langle \text{ИГ}_{\text{м,ед,им}}, S_{\text{м,ед,им}} \rangle$, ставим после *мальчик* еще одну скобку с пометкой «$\text{ИГ}_{\text{м,ед,им}}$», убираем из магазина верхнее предсказание и переходим к очередному слову.

После 2-го шага наша фраза имеет вид

0	1	2	2	2
(	(*Маленький*	(*мальчик*	)	)
ПРЕДЛ	$\text{ИГ}_{\text{м, ед, им}}$	$\text{ИГ}_{\text{м, ед, им}}$	$\text{ИГ}_{\text{м, ед, им}}$	$\text{ИГ}_{\text{м, ед, им}}$

на далеком от города... (цифра над скобкой означает номер шага, на котором эта скобка появилась), а магазин имеет вид 2б на стр. 146.

3-й шаг. Рабочее слово — *на*, рабочая пара — ⟨$ГГ_{м,ед,3}$, $Pr_{предл}$⟩; она есть в таблице А (строка 9). Открываем левую скобку глагольной группы и берем одно из имеющихся в табл. А СП этой пары: $ИГ_{м,ед,предл}$ + $ГГ_{м,ед,3}$, которое помещается в магазин (табл. 3).

4-й шаг. Рабочее слово — *далеком*, рабочая пара — ⟨$ИГ_{м,ед,предл}$, $А_{м,ед,предл}$⟩. Найдя ее в таблице А, открываем перед *далеком* скобку с пометкой «$ИГ_{м,ед,предл}$» и в качестве очередного СП выбираем ПГ+$ИГ_{м,ед,предл}$ (строка 3); магазин приобретает вид табл. 4.

5-й шаг. Рабочее слово — *от*, рабочая пара — ⟨ПГ, $Pr_{род}$⟩. Поскольку она есть в табл. А, ставим перед *от* левую скобку («ПГ») и помещаем в магазин СП данной пары (строка 7); теперь магазин содержит предсказания (сверху вниз): $ИГ_{м,ед,род}$ + ПГ + $ИГ_{м,ед,предл}$ + $ИГ_{м,ед,предл}$+$ГГ_{м,ед,3}$+ $ГГ_{м,ед,3}$ + + ПРЕДЛ (табл. 5).

6-й шаг. Рабочее слово — *города*, рабочая пара — ⟨$ИГ_{м,ед,род}$, $S_{м,ед,род}$⟩. Эта пара есть в табл. А; открываем перед *города* скобку («$ИГ_{м,ед,род}$»), но вместо пути 2а, как на шагах 3—5, выбираем путь 2б. Найдя рабочую пару в табл. Ω, закрываем скобку после *города* и убираем рабочее предсказание $ИГ_{м,ед,род}$ из магазина (табл. 6а); затем повторяем операцию 2б еще раз: новая рабочая пара ⟨ПГ, $S_{м,ед,род}$⟩ содержится в табл. Ω (предложная группа может оканчиваться существительным), поэтому ставим после *города* еще одну правую скобку, теперь с пометкой ПГ, и убираем рабочее СП из магазина (табл. 6б), после чего переходим к очередному слову.



Оставшиеся шаги (7—10) читатель может выполнить сам; чтобы облегчить ему эту задачу, мы приведем состояния магазина после каждого из шагов 7—10 (табл. 7а — 10д) и фразу в том виде, какой она должна иметь после завершения анализа (цифра над скобкой — это номер шага, на котором

появилась данная скобка; цифра под скобкой — сокращенное обозначение типа группы):

0 1 2 2 2 3 4 5 6 6 6
((*маленький* (*мальчик*)) (*на* (*далеком* (*от* (*города*))
1 2 2 2 2 3 4 5 6 6 5

7 7 7 8 9 9 10 10 10 10 10
(*полустанке*)) (*ждал* (*приезда* (*родителей*)))))
4 4 4 3 6 7 7 6 3 3 1

Обозначения типов групп:

1 — ПРЕДЛ 4 — $\text{ИГ}_{\text{м, ед, предл}}$ 6 — $\text{ИГ}_{\text{м, ед, род}}$
2 — $\text{ИГ}_{\text{м,ед,им}}$ 5 — ПГ 7 — $\text{ИГ}_{\text{м,мн,род}}$
3 — $\text{ГГ}_{\text{м, ед, з}}$

Состояния магазина:

0	1	2а
$\text{ИГ}_{\text{м, ед, им}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ	$\text{ИГ}_{\text{м, ед, им}}$ $\text{ИГ}_{\text{м. ед, им}}$ $\text{ГГ}_{\text{м. ед, з}}$ ПРЕДЛ	$\text{ИГ}_{\text{м, ед, им}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ
2б	**3**	**4**
$\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ	$\text{ИГ}_{\text{м, ед, предл}}$ $\text{ГГ}_{\text{м, ед, з}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ	ПГ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ГГ}_{\text{м, ед, з}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ
5	**6а**	**6б**
$\text{ИГ}_{\text{м, ед, род}}$ ПГ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ГГ}_{\text{м, ед, з}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ	ПГ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ГГ}_{\text{м, ед, з}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ	$\text{ИГ}_{\text{м, ед, предл}}$ $\text{ИГ}_{\text{м, ед, предл}}$ $\text{ГГ}_{\text{м, ед, з}}$ $\text{ГГ}_{\text{м, ед, з}}$ ПРЕДЛ

7а	7б	8
$ИГ_{м, ед, предл}$ $ГГ_{м, ед, 3}$ $ГГ_{м, ед, 3}$ ПРЕДЛ	$ГГ_{м, ед, 3}$ $ГГ_{м, ед, 3}$ ПРЕДЛ	$ИГ_{м, ед, род}$ $ГГ_{м, ед, 3}$ $ГГ_{м, ед, 3}$ ПРЕДЛ
9	**10а**	**10б**
$ИГ_{м, мн, род}$ $ИГ_{м, ед, род}$ $ГГ_{м, ед, 3}$ $ГГ_{м, ед, 3}$ ПРЕДЛ	$ИГ_{м, ед, род}$ $ГГ_{м, ед, 3}$ $ГГ_{м, ед, 3}$ ПРЕДЛ	$ГГ_{м, ед, 3}$ $ГГ_{м, ед, 3}$ ПРЕДЛ
10в	**10г**	**10д**
$ГГ_{м, ед, 3}$ ПРЕДЛ	ПРЕДЛ	

Таким образом, получается вполне обычное представление НС-структуры, правда, с одной несущественной особенностью: атомарные составляющие (индивидуальные словоформы), начинающие более крупные составляющие, не заключаются в скобки. Это связано со спецификой работы П-анализатора. а именно с тем, что он обрабатывает фразу только слева направо.

Заметим, что наша фраза имеет и другую синтаксически правильную интерпретацию: ((*Маленький* (*мальчик* (*на* (*далеком* (*от* (*города*)) (*полустанке*))))) (*ждал* (*приезда* (*родителей*)))), т. е. группа *на... полустанке* относится не к группе сказуемого, а к группе подлежащего (*мальчик на полустанке*, а не *ждал на полустанке*). Хотя такая интерпретация семантически вряд ли приемлема, синтаксически она вполне законна и должна выдаваться П-анализатором. В нашем примере анализа мы могли бы получить эту интерпретацию, если бы на 2-м шаге мы выбрали путь 2а, а не 2б, и при этом взяли бы из



таблицы А предсказание ПГ (строка 4). Рекомендуем читателю провести анализ по этому пути самостоятельно.

С формальной точки зрения П-анализатор представляет собой частный случай так называемого автомата с магазинной памятью (М-автомата). Дать точное определение М-автомата в рамках настоящего изложения, не рассчитанного на читателя, знакомого с теорией алгоритмов, слишком сложно, поскольку для его понимания фактически требуется знание машин Тьюринга. Для подготовленного читателя это определение приводится ниже мелким шрифтом.

Недетерминированная машина Тьюринга *) называется *М-автоматом*, если она удовлетворяет следующим условиям:

1. Машина имеет три ленты (называемые соответственно входной, рабочей и выходной) и три головки, по одной на каждой ленте. Все три ленты ограничены слева и неограничены справа.

2. Машина имеет три внешних алфавита: в х о д н о й — для входной ленты (содержательно — основной словарь), р а б о ч и й — для рабочей ленты (содержательно — набор синтаксических предсказаний + «пустой символ») и в ы х о д н о й (содержательно — набор левых и правых скобок, помеченных символами типов составляющих, + «пустой символ»).

3. Машина может производить следующие элементарные операции:

а) На входной ленте — только читать символ, записанный в обозреваемой ячейке, и сдвигать головку на одну ячейку в п р а в о.

б) На рабочей ленте — либо писать в обозреваемой ячейке (непустой) символ и сдвигать головку на одну ячейку в п р а в о, либо сдвигать головку на одну ячейку в л е в о и стирать записанный в ней символ. (Таким образом, рабочая лента всегда имеет вид, показанный на стр. 149.)

Из этого рисунка видно, что рабочая лента есть не что иное, как магазин.

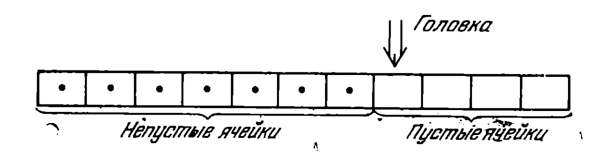
в) На выходной ленте — только записывать символ и сдвигать головку на одну ячейку в п р а в о. (В результате работы машины на выходной ленте должно выписываться

*) Недетерминированная машина Тьюринга отличается от обычной машины Тьюринга только тем, что в ее программе могут быть различные команды с тождественными левыми частями.

скобочное представление системы составляющих для анализируемой фразы.)

г) На каждом шаге машина работает только на одной ленте.

д) В начальный момент все три головки находятся в крайних левых ячейках соответствующих лент, рабочая и



выходная ленты пусты. Работа машины считается законченной, если головка на входной ленте дошла до первой пустой ячейки, а рабочая лента в этот момент пуста.

Читатель, знакомый с машинами Тьюринга, без труда сообразит, что П-анализатор соответствует М-автомату, на который наложено следующее дополнительное ограничение: перед каждой записью символа на рабочую ленту машина должна читать символ на входной ленте. (Содержательно это ограничение объясняется тем, что помещение синтаксического предсказания в магазин П-анализатора всегда непосредственно обусловливается прочтением очередного слова фразы.) Заметим еще, что М-автомату разрешается писать за один шаг на рабочей ленте только один символ, тогда как П-анализатор иногда пишет сразу несколько символов; очевидно, что это различие не принципиально.

Для каждого М-автомата можно рассматривать множество, состоящее из всех тех и только тех цепочек (отрезков текста), которые он способен проанализировать, — язык, определяемый М-автоматом, или М-язык (ср. понятие К-языка, стр. 132). Доказано, что

$\mathbf{T}_{2.3}$ а) для каждого М-автомата можно построить эквивалентную ему КС-грамматику;

$\mathbf{T}_{2.4}$ б) для каждой КС-грамматики можно построить эквивалентный ей М-автомат, т. е. М-автомат, определяющий тот же язык.

Таким образом, класс М-языков совпадает с классом КС-языков (а тем самым в силу теоремы $T_{2.2}$, стр. 133, с классом К-языков).

О значении этого факта уже говорилось выше, стр. 93—94.

К вопросу о классификации грамматик

Закончив обзор основных типов грамматик, мы можем теперь высказать следующее соображение относительно их классификации. В литературе по математической лингвистике грамматики всегда подразделяют на порождающие и распознающие; этой принятой классификации придерживались и авторы данной книги, хотя она представляется им в значительной мере условной.

В самом деле, прежде всего легко видеть, что порождающие грамматики наиболее важного типа, а именно, неукорачивающие (и, в частности, НС-грамматики), могут быть использованы также и для распознавания, т. е. для отличения грамматически правильных (выводимых) предложений от неправильных (невыводимых). Это фактически было доказано на стр. 50—53. Для произвольных грамматик такая ситуация не имеет места: существуют грамматики, для которых алгоритм распознавания выводимости цепочек невозможен (ср. первую сноску на стр. 50). Однако суть дела в том, что порождению естественно противопоставляется не распознавание, а, так сказать, «допускание». Именно, естественно говорить, что некоторая грамматика Γ допускает язык L, если Γ дает процедуру, способную для любой цепочки x, принадлежащей L, рано или поздно установить это (т. е. установить, что $x \in L$); если же x не принадлежит L, то от этой процедуры ничего не требуется: она может либо обнаружить, что $x \notin L$, либо не дать вообще никакого результата, т. е. продолжаться вечно. (Напомним, что от распознающей процедуры требуется больше: она должна давать результат в любом случае — положи-



тельный, если $x \in L$, и отрицательный, если $x \notin L$.)

Если вместо распознавания рассматривать «допускание», то тогда все порождающие грамматики могут трактоваться и как допускающие. При этом допускающая процедура состоит просто в том, что правила грамматики применяются к данной цепочке «наоборот» — справа налево: в цепочке отыскивается вхождение правой части некоторого правила и заменяется левой частью, и этот процесс продолжается, пока можно. Допускаемыми цепочками будут в точности те, которые могут быть свернуты указанным процессом к начальному символу; ясно, что это как раз те самые цепочки, которые при «обычном» использовании грамматики выводятся из начального символа.

С другой стороны, распознающие грамматики (которые заведомо являются и допускающими) могут использоваться и для порождения. Так, чтобы с помощью категориальных грамматик можно было порождать цепочки, достаточно переформулировать правила сокращения как правила развертывания (т. е. фактически просто прочитать их наоборот): 1) всякую категорию Ψ можно развернуть в $\Phi[\Phi\backslash\Psi]$, где Φ — произвольная категория (левое развертывание), 2) всякую категорию Φ можно развернуть в $[\Phi/\Psi]\Psi$, где Ψ — произвольная категория (правое развертывание). Легко сообразить (читатель, по-видимому, без труда сделает это сам), как будет тогда осуществляться процесс порождения. Ср. стр. 133.



Что же касается М-автоматов, то излагать способ использования их для порождения мы не будем; для нас здесь важно лишь, что это, безусловно, можно сделать.

Таким образом, формальные грамматики по существу нейтральны по отношению к порождению и допусканию (а также распознаванию, если оно возможно). Можно полагать, что было бы целесообразно говорить не о порождающих и распознающих (допускающих) грамматиках, а просто о граммати-

ках, рассматривая аспект «направление применения» (для порождения или допускания) *отдельно* от определения самих грамматик. Тогда конкретная грамматика будет трактоваться как «грамматика такого-то класса в таком-то аспекте», например, «допускающая НС-грамматика», «порождающий М-автомат» и т. п.

Обычное деление грамматик на порождающие и распознающие имеет естественное историческое объяснение. Те грамматики, которые называют порождающими (соответственно распознающими), разрабатывались с целью использовать их как раз для порождения (распознавания). Однако, как мы только что видели, независимо от цели создания грамматики она может использоваться «в обе стороны». Поэтому противопоставление «порождающие/ / распознающие» представляется недостаточно отражающим существо дела.

Более глубоким, как нам кажется, является *различие в способе задания информации о синтаксической сочетаемости*: эта информация содержится либо в правилах (любого вида), либо в словаре, т. е. в синтаксических кодах отдельных слов (в последнем случае правила тоже нужны, но их сравнительно немного и притом они не содержат никакой информации о конкретном языке; ср. два правила сокращения К-грамматик). На этом различии может основываться подразделение грамматик на два главных класса: грамматики с правилами (rule grammars; к сожалению, невозможен термин «правильные грамматики») и грамматики без правил (ruleless grammars, «бесправильные грамматики»). К первому классу относятся грамматики Хомского и М-автоматы (для последних роль правил играют синтаксические таблицы A и Ω), ко второму — К-грамматики (ср. стр. 123).

Ни тот, ни другой класс не исчерпываются названными типами грамматик; новых типов, однако, мы привлекать здесь не будем, а ограничимся теми, которые рассматривались в §§ 2—6.

В заключение заметим, что речь идет здесь не просто о новой классификации, но скорее о новом аспекте рассмотрения грамматик. Обычно в понятие грамматики неявно включается и способ ее использования, хотя в формальных определениях грамматик он не фигурирует; предлагается же, как уже было сказано выше, полностью отделить этот способ от самого понятия грамматики. Вместе с тем на первый план выдвигается способ подачи информации о конкретном языке; эта особенность грамматик отражается в их определениях.

«Логический анализ языка»

Все рассматривавшиеся нами до сих пор модели языка — формальные грамматики — фактически имели дело только с одним, хотя и исключительно важным, аспектом естественного языка: с синтаксисом в широком смысле этого слова, т. е. с правилами комбинирования некоторых исходных единиц друг с другом в тексте. Синтаксис в широком смысле включает в себя как синтаксис в традиционном лингвистическом понимании, т. е. законы построения



словосочетаний из слов, предложений из словосочетаний и т. п., так и морфологию, т. е. законы построения словоформ из морф. Разумеется, различение морфологии — «действует внутри слова» — и собственно синтаксиса — «действует вне слова» — вполне целесообразно и даже необходимо. Однако с более общей точки зрения их удобно объединять вместе и называть «синтаксисом в широком смысле», что здесь и делается.

Иначе говоря, формальные грамматики задают, как уже указывалось выше, стр. 106, только п р а в и л а о б р а з о в а н и я языковых выражений.

Однако у языка есть и другой, не менее важный аспект: семантика, для описания которой необходимо ввести в рассмотрение правила перехода от одних выражений языка к некоторым другим выражениям, несущим ту же информацию (имеющим тот же смысл), т. е. п р а в и л а п р е о б р а з о в а н и я

языковых выражений, ср. стр. 107. Смысл языкового выражения рассматривается здесь как то общее, что имеется у данного выражения и всех других синонимичных ему выражений того же или какого-либо другого (не обязательно естественного) языка. Иными словами, смысл есть инвариант синонимичных преобразований (ср. положение Р. О. Якобсона о том, что означающее — это то, что воспринимается, а означаемое — это то, что переводится). При таком подходе описать семантику некоторого естественного языка L означает: 1) указать другой язык L', на который будут переводиться выражения языка L, и 2) задать правила перевода любых выражений с L на L' и обратно. Для практических целей, в частности для изучения семантики иностранного языка, в качестве L' обычно берется какой-либо другой естественный язык, например родной язык учащегося. Если же речь идет об описании семантики языка в научных теоретических целях, то L', очевидно, не может быть естественным языком: научное описание предполагает максимальную логическую четкость, эксплицитность и однозначность, а в этом отношении ни у одного естественного языка нет существенных преимуществ перед другими. Поэтому для научного описания семантики естественного языка язык L' должен быть сконструирован специально и, как полагают авторы, удовлетворять по крайней мере следующим четырем требованиям:

1. Все элементы L' должны находиться в одно-однозначном соответствии с обозначаемыми сущностями — исключается не только явная омонимия типа *лук*$_1$ (оружие) — *лук*$_2$ (овощ) или *глава*$_1$ (в книге) — *глава*$_2$ (семейства), но и полисемия, как, например, в *допускать*: *допускать*$_1$ (к документам) — — *допускать*$_2$ (много решений, о задаче). Для всех подобных случаев в L' необходимо иметь разные элементы. С другой стороны, скажем, отношение «субъект — действие», которое в естественных языках может оформляться многими способами (ср. *шлейхеровская формулировка* — *формулировка Шлейхера* — *формулировка Шлейхером* [*закона*] —

Шлейхер формулирует [*закон*] и т. п.), в L' должно иметь одно обозначение.

2. L' должен иметь достаточно возможностей (быть достаточно богатым) для описания смыслового содержания любых выражений языка L или хотя бы его существенного фрагмента. При этом L' не должен иметь обязательных смысловых категорий — таких, употребление которых навязывалось бы его собственными правилами, наподобие числа существительных и времени глаголов в русском языке (по-русски невозможно построить фразу, где существительные были бы вообще «без числа», а глаголы — «без времени») или кванторов в языке логики предикатов (при стандартном употреблении этого языка переменные в правильных формулах обязательно должны быть связаны кванторами *)). Наличие в L' своих обязательных категорий вносило бы, по-видимому, искажения в картину семантики изучаемого языка. Разумеется, L' должен иметь средства для выражения всех понятий, связанных с числом или временем, но употребление этих средств не должно быть обязательным, подобно тому как в русском языке можно выражать размер или цвет предметов, но делать это не обязательно.



3. L' должен обладать достаточной, но не чрезмерной «разрешающей способностью»: в нем должны быть выразимы все те смысловые различия, имеющиеся в L, которые интуитивно представляются нам существенными; в то же время L' не должен различать больше, чем различается в L: ведь L' служит не для усовершенствования языка L, а для его описания. Заметим, что при изучении семантики естественных языков исследователи — по большей части философы и логики — обычно концентрируют свое внимание на том, как бы не упустить какие-либо различия, забывая, что не менее опасно вносить

*) Фактически часто допускаются формулы, содержащие свободные, т. е. не связанные квантором, переменные; однако в таких случаях, как правило, подразумевается квантор всеобщности (т. е. используется «нулевое» обозначение квантора).

лишние различия, идущие от логики, а не от языкового материала.

4. Язык L' должен быть совершенно формальным, т. е. он должен быть задан каким-либо четким и абсолютно явным способом, например, с помощью порождающей грамматики.



Язык, удовлетворяющий перечисленным требованиям, можно назвать семантическим языком. Таким образом, модель естественного языка, ориентированная на его семантический аспект, есть некий семантический язык плюс правила перевода с этого семантического языка на естественный и обратно.

Построение конкретных семантических языков и правил перевода вряд ли следует относить к самой математической лингвистике, как не относится к ней, например, и разработка конкретных порождающих грамматик для конкретных языков: математическая лингвистика занимается лишь общей теорией грамматик, изучением их абстрактных свойств. Аналогично этому в математическую лингвистику должна была бы входить только общая теория семантических языков.

Однако такой теории в настоящее время еще не существует. Более того, хотя ее разработка представляется чрезвычайно важной как в теоретическом, так и в прикладном плане, авторам неизвестны работы, ведущиеся в соответствующем направлении. Не существует пока что и готовых семантических языков, во всяком случае — достаточно полных. Однако можно указать ряд частичных семантических языков, построенных для моделирования весьма узких и очень специальных фрагментов семантики (см. ниже, стр. 161 —162). Поскольку эти языки являются базой и для разработки полных семантических языков, и для построения теории этих последних, стоит, по-видимому, сказать о них несколько слов.

Классическим примером частичного семантического языка является язык логики предикатов. Последующие замечания рассчитаны на читателя, уже

знакомого с этим языком *), поскольку описывать здесь этот язык было бы нецелесообразно. Однако ввиду исключительной важности этого языка для общей культуры мышления, особенно для построений, связанных с естественным и другими искусственными языками (именно язык логики предикатов фактически лежит в основе известных авторам частичных семантических языков), мы настоятельно рекомендуем читателю, еще не овладевшему языком логики предикатов, сделать это. По существу это совсем несложно и не требует особой математической подготовки; однако возможны затруднения в связи с отсутствием элементарных и доступных руководств (на русском языке). Впрочем, можно посоветовать обратиться к Гильберт — Аккерман 1947, Тарский 1948 или Mendelson 1964.

Язык логики предикатов — основной аппарат математической логики — обладает высокой степенью формализации и очень хорошо изучен. Сам по себе он предназначен для описания весьма ограниченной части семантики, а именно той, которая имеет дело с истинностью или ложностью утверждений. Тем не менее его элементы — логические связки, кванторы и, в особенности, сами предикаты — допускают более широкое использование. Этим и объясняется тот факт, что большинство попыток формального описания семантики естественных языков связано с применением языка логики предикатов или хотя бы его составных частей.

Подобные попытки можно разделить на два типа. К первому относятся исследования, в которых естественные языки сопоставляются с языком логики предикатов с целью обнаружить в них единицы и категории, аналогичные единицам и категориям этого последнего, т. е. выяснить, как выражаются в естественных языках логические категории. Эти работы обычно и объединяют под общим названием

---

*) Здесь имеется в виду знакомство именно лишь с языком логики предикатов, а не с ней самой: знания каких-либо теорем математической логики не требуется.

«логический анализ языка». К сожалению, исследования в области логического анализа языка, очень важные и перспективные по своему направлению, представляют собой, как правило, отдельные наблюдения и соображения, нередко очень глубокие, но не образующие единой связной теории. (Это, возможно, объясняется тем, что логическим анализом языка до последнего времени занимались по большей части не лингвисты, а логики, для которых естественные языки являются не столько самостоятельным объектом описания, сколько иллюстративным материалом.) Поэтому мы вынуждены ограничиться простым упоминанием о нескольких работах подобного рода, не претендуя ни в малейшей степени на полноту охвата хотя бы важнейших из них.

Еще в 1940 году Б. Рассел обратил внимание на то, что в естественных языках предлоги и глаголы по своей семантической природе совпадают, а именно, и те и другие суть имена предикатов; например, предлог *до* и глагол *предшествовать*, при всем различии их грамматических свойств, означают одно и то же — они выражают двухместный предикат «А предшествует В» (Russel 1940, стр. 124)*). Из этого вытекает, в частности, очень важная мысль (впоследствии развитая в целом ряде работ), что тождество/различие смыслов слов естественного языка может быть определено путем сопоставления им точно определенных предикатов.

Особое место среди работ по логическому анализу языка занимает раздел «Анализ разговорного языка» книги Г. Рейхенбаха (Reichenbach 1960, стр. 251—354). Здесь развивается идея классификации слов, исходящей из их логико-семантической природы, а не из их формально-грамматических свойств, т. е. классификации, отличной от классификации по частям речи. Примерами логико-семантических классов являются класс одноместных предикатов

*) Любопытно, что в ряде естественных языков, например во вьетнамском, предлог и глагол грамматически принадлежат к одной категории: вьетн. *ở* — предлог '*в*' и глагол '*находиться в*'.

(глаголы вроде *спать*, прилагательные вроде *умный*, существительные вроде *плакса*), класс двухместных предикатов (*любить*, *читать*...; *похожий*, *приятный*... ...; *жена*, *начальник*...; *поиск*, *исполнение*...; *над*, *позади*...), класс предметных констант (*нос*, *дерево*, *пушка*), класс кванторных слов (*все*, *один*, *везде*, *иногда*, *нечто*), класс логических связок (*и*, *или*, *а*, *но*, *если*... *то*) и т. п. Логико-семантическая классификация слов имеет глубокий лингвистический смысл; в частности, принадлежность слова к тому или иному классу существенно сказывается на его синтаксических свойствах. Кроме того, Рейхенбах рассмотрел еще целый ряд важных вопросов, из которых мы назовем только попытку анализа категории артикля с помощью понятий ι-оператора ($\approx$ «тот, который...») и ε-оператора ($\approx$ «такой, который...»), а также набросок универсальной схемы глагольных времен. Данная работа Рейхенбаха стала классической, и большинство последующих работ по логическому анализу языка в той или иной степени на нее опираются.

Постановка задачи логического анализа языка и некоторые интересные примеры имеются в статье Quine 1961 (замечания о парных союзах как аналогах скобок, о словах типа *every* — *any*, о личных местоимениях — аналогах предметных переменных и т. д.). См. также Quine 1960.

Следует отметить серию работ Элинор К. Чарни (например, Charney 1961, 1962), в которых исследуется логическое строение значений английских союзов (типа *if*, *unless*) и кванторных слов (*all*, *every*, *any*...). Среди других вопросов здесь рассматривается, в частности, вопрос о синонимии таких предложений, как *To invite all women and no men is to make a dull party* и *To invite only women and no men is to make a dull party*, означающих в точности одно и то же ('Если пригласить только женщин и вовсе не пригласить мужчин, вечеринка будет скучной'), хотя различающие их слова *all* 'все' и *only* 'только' отнюдь не являются синонимами. Результаты указанных работ подытожены в недавно вышедшей

книге (Charney 1966), где излагаются также некоторые общие соображения относительно изучения семантики логическими средствами.

В работе К. Демана (Döhmann 1966) дается обзор способов, которыми выражаются все двухместные функции алгебры логики (конъюнкция: $A \& B$, дизъюнкция: $A \vee B$, строгая дизъюнкция: $(A \vee B) \& \neg (A \& B)$, импликация: $A \supset B$ и т. д., всего 16) в самых разных естественных языках. Так, конъюнкция может быть выражена простым сопоножением (кит. *ма лю* 'лошадь и осел', букв 'лошадь осел'), простым союзом (*и*, *and*, *et*), парным союзом (*как... так и*), предлогом (*Петя* **с** *Машей пришли...*), постпозитивной частицей (лат. *Senatus populus***que** *romanus* 'Сенат и римский народ') и т. п. Обратный подход применен Е. В. Падучевой в статье Падучева 1964в, где выясняется, что русским союзом *или* выражаются три разные логические функции: строгая дизъюнкция (*Твоя книга лежит в шкафу или на столе* *)), нестрогая дизъюнкция (в контексте явной или «скрытой» импликации: *Если у меня заболит горло или повысится температура, то я не поеду кататься на лыжах*; *Студенты, выступавшие с докладом или подавшие письменный отчет, освобождаются от экзамена*) и конъюнкция ($x^2 > 0$ *при* $x > 0$ *или при* $x < 0$).

В связи с проблемой перевода «язык логики предикатов — русский язык» Е. В. Падучева рассмотрела также (Падучева 1964а и Падучева 1964г) вопрос о средствах выражения в русском языке скобок, ограничивающих области действия логических связок; оказалось, что русский язык позволяет обеспечить однозначность понимания с точки зрения расстановки таких скобок: роль левых скобок играют части парных союзов, а без правых скобок можно обойтись [ср. $(A \& B) \vee C$ = *или* A *и* B, *или* C; $A \& (B \vee C)$ = *как* A, *так и* B *или* C].

---

*) Ср. *Твои книги лежат в шкафу или на столе*, где из-за мн. числа слова *книги* союз *или* имеет значение нестрогой дизъюнкции.

Пожалуй, наиболее систематический и обобщающий обзор особенностей естественных языков с логической точки зрения можно найти в статье У. Вейнрейха (Weinreich 1963); однако изложение ее содержания завело бы нас слишком далеко.

Теперь мы коснемся попыток другого типа (напомним, что речь идет о двух типах попыток формального описания семантики, см. выше, стр. 157) — а именно, попыток, связанных с конструированием на базе языка логики предикатов более широких семантических языков. Поскольку описать даже простой язык достаточно кратко невозможно, нам придется отказаться от всяких содержательных разъяснений по существу перечисляемых здесь работ; таким образом, то, что говорится ниже, следует рассматривать просто как библиографическую справку.

Наиболее известны так называемые информационные языки, создаваемые для конкретных областей науки и предназначаемые для записи сведений из этих областей в такой простой, явной и однозначной форме, которая была бы удобна для машинной обработки (автоматический поиск нужной информации, автоматический вывод следствий из данных посылок, автоматическое реферирование и т. д.). В качестве примера можно сослаться на информационный язык для элементарной геометрии, разрабатывавшийся в ВИНИТИ АН СССР (см. Кузнецов и др. 1961).

Любопытным примером семантического языка (приобретшим благодаря его необычному назначению широкую известность — не в том смысле, что многие знают его, а в том, что многие знают о нем) является построенный Г. Фройденталем «Линкос» (Lingua Cosmica), предназначаемый для общения с внеземными цивилизациями (Freudenthal 1960).

Наконец, имеется ряд семантических языков, разработанных для целей автоматического перевода, в процессе которого семантический язык должен выступать в роли языка-посредника. Здесь мы

упомянем семантические языки Кэмбриджского лингвистического кружка (Мастерман 1964 и Паркер-Роудс 1964), а также язык «СМ-1» для представления математических текстов (Гладкий и др. 1961; в отличие от только что упомянутого информационного языка для геометрии «СМ-1» должен служить для изображения не самого математического содержания, а фраз естественного языка, составляющих математический текст; поэтому он не годится, например, для автоматического вывода следствий). Фактически семантический язык («предикатная запись с использованием семантических множителей») предложен в работе Жолковский — Леонтьева — Мартемьянов 1961; его развитие и существенное обогащение привели к новому языку, описанному в Жолковский — Мельчук 1967 («лексико-синтаксические структуры и система перифразирования»). Заметим, что явного формального задания соответствующих языков в двух последних работах не содержится.

В заключение этого раздела еще раз отметим (ср. стр. 156), что все изложенное в нем фактически не входит в математическую лингвистику, так что этот раздел по содержанию (да и по стилю, как читатель, вероятно, уже заметил) выпадает из рамок нашей книги. Тем не менее мы сочли целесообразным включить его в книгу, чтобы, так сказать, хотя бы «застолбить» соответствующий участок, т. е., во-первых, подчеркнуть необходимость «семантического раздела» в математической лингвистике и, во-вторых, указать тот материал, на основе которого такой раздел, по-видимому, должен быть создан.

Б. Моделирование лингвистических исследований

До сих пор мы занимались исключительно формализацией описания языка как такового, т. е. моделированием системы, определяющей речевую деятельность. В самом деле, рассмотренные выше формальные грамматики всех типов представляют

собой модели языка: их «поведение» (порождение или распознавание текстов) в некоторых существенных аспектах сходно с поведением людей, общающихся при помощи языка. Строя эти модели, мы широко пользовались нашими интуитивными представлениями об общих свойствах человеческого языка, практическим знанием конкретных языков (в наших примерах — русского), а также целым рядом принятых и вполне обычных в лингвистике (а по большей части и в школьной грамматике) понятий и категорий таких, как «существительное», «глагол», «род», «падеж», «именная группа» и т. п. При этом мы не задавались вопросом, откуда у нас те или иные представления о языке, как именно мы осуществляем научное изучение языков, почему выбраны такие категории (а не какие-нибудь другие) и что они собой представляют и т. д. Мы имели полное право поступать так, поскольку нас интересовали только сами формальные грамматики безотносительно к тому, как мы к ним приходим. Однако этот вопрос вполне закономерен и, более того, чрезвычайно интересен. Чтобы ответить на него, надо создавать формальные модели тех процедур и тех средств, которыми мы фактически пользовались при построении грамматик, никак не уточняя их. Другими словами, надо создавать формальные модели деятельности лингвиста (или вообще человека, овладевающего языком). Результатом работы системы таких моделей в конце концов должны быть модели языка, т. е. формальная модель лингвистического исследования есть модель процесса построения моделей языка. О подобных моделях и пойдет речь ниже.

Мы не будем давать здесь сколько-нибудь полный обзор существующих или возможных моделей лингвистического исследования, а ограничимся краткой характеристикой двух конкретных моделей. Первая относится к аспекту анализа (формализации) основных понятий и категорий лингвистики, а вторая — к аспекту формализации самих процедур исследования. Разумеется, оба эти аспекта

тесно связаны: ведь процедура, ведущая к выделению некоторого класса объектов, может рассматриваться как формальное определение этого класса — конструктивное определение. Тем не менее их целесообразно разграничивать: во-первых, не всякое формальное определение задает конструктивную процедуру, а во-вторых, указанные аспекты различаются точкой зрения, логическим акцентом — в случае первого аспекта преимущественное внимание обращается на само анализируемое понятие, в случае второго — на процедуру, ведущую к обнаружению соответствующих объектов.

В настоящее время имеется ряд работ, в которых делаются попытки формально определить различные традиционные лингвистические понятия — часть речи, падеж, род, фонему и т. п.*) (Зализняк 1967, 1969, Ревзин 1967, Успенский 1964, Маркус 1964, Marcus 1963). Мы выберем для примера те из них, которые посвящены понятию составляющей. Ясно, что это понятие совершенно необходимо для описания языка. Так называемая структурная лингвистика широко использует его в явном виде; традиционное же языковедение, хотя практически и не пользуется этим термином, по существу имеет дело с более или менее аналогичными понятиями: словосочетание, синтагма (в смысле В. В. Виноградова), синтаксическая группа, группа члена предложения и т. п., причем все эти понятия не определены с достаточной четкостью и нередко понимаются по-разному даже учеными, принадлежащими к одной школе. Это делает построение формального аналога для понятия составляющей важной и интересной задачей (тем более в рамках настоящего изложения, где это понятие уже неоднократно использовалось — ср., например, стр. 61 и сл.).

Для облегчения формальных рассмотрений еще раз приведем примеры составляющих. Во фразе

*) Большинство подобных работ советских и некоторых зарубежных исследователей восходит, в той или иной степени, к пионерской работе О. С. Кулагиной (Кулагина 1958).

Снаружи вьюга мечется и все заносит в лоск имеется 14 составляющих:

1) вся фраза целиком;
2) *вьюга мечется и все заносит в лоск*;
3) *мечется и все заносит в лоск*;
4) *все заносит в лоск*;
5) *все заносит*;
6) *в лоск*;

7—14) каждое слово, взятое само по себе.

Два важных замечания:

1. Разложение на составляющие не обязательно должно быть единственным даже в синтаксически однозначной фразе. Например, ...*на черном дне твоих зеленых глаз* может быть разложено на составляющие разными способами:

((*на* (*черном дне*)) (*твоих* (*зеленых глаз*))), или
(*на* ((*черном дне*) (*твоих* (*зеленых глаз*)))), или
(*на* (*черном* (*дне* (*твоих* (*зеленых глаз*))))), или
((*на* (*черном дне*)) ((*твоих зеленых*) *глаз*)),

причем это еще не все возможные разложения. (По-видимому, читателю будет полезно самостоятельно продолжить этот список.)

2. Слова, способные образовывать составляющую, не обязательно образуют ее в любой фразе. Если иметь в виду «естественную» систему составляющих, то, например, сочетание *дело мастера* является составляющей во фразе *Это, безусловно, дело мастера*, и не является таковой во фразе *Дело мастера боится*; аналогично сочетание *списки номеров, которые хранятся в памяти* является составляющей во фразе *Рассмотрим списки номеров, которые хранятся в памяти* и не является составляющей во фразе *Лишим списки номеров, которые хранятся в памяти*.

Теперь мы можем перейти к формальному определению составляющей. Точнее, будет определено понятие к о н ф и г у р а ц и и, или потенциальной составляющей. Содержательно, конфигурация— это такой отрезок текста (цепочка), который может быть составляющей хотя бы в одной фразе

данного языка. Так, отрезки *дело мастера* или *списки номеров, которые хранятся в памяти* суть конфигурации, хотя они и не всегда составляющие (см. выше) *).

Определим сначала понятия замещаемости и взаимозамещаемости. Пусть имеется некоторый язык L со словарем V (т. е. некоторое множество цепочек из входящих в V символов). Рассмотрим две цепочки x и y, состоящие из символов словаря V (NB: x и y не обязательно принадлежат L). Цепочка x замещаема на y в языке L [обозначение: $x \Rightarrow y$ (L)], если для любых двух цепочек p и q из $pxq \in L$ следует $pyq \in L$ (p и q не обязательно принадлежат L; в частном случае они могут быть пустыми). Содержательно это означает, что в любой правильной фразе цепочку x можно заменить на y, не нарушая правильности; обратное, вообще говоря, неверно. Например, словоформа *чай* замещаема словоформой *кофе*, но не наоборот (*пачка кофе*, но не* *пачка чай*); *напевы* замещаемо на *забытые напевы*, но не наоборот (*давно забытые напевы*, но не **давно напевы*).

Цепочка x взаимозамещаема с y в языке L [обозначение: $x \Leftrightarrow y$ (L)], если $x \Rightarrow y$ (L) и $y \Rightarrow x$ (L). Примеры: *песком* и *гравием*, *исключительно важных* и *важных* и т. п.

Понятие конфигурации определяется индуктивно: сначала мы определим конфигурацию 1-го ранга (базис индукции), а затем через нее — конфигурации высших рангов (индукционный шаг).

а) Базис индукции. Конфигурация 1-го ранга языка L — это цепочка, состоящая не менее чем из двух символов и взаимозамещаемая с одним символом, который называется результирующей данной конфигурации.

NB: 1. Результирующая не обязательно должна быть одним из символов, входящих в конфигурацию.

---

*) Обращаем внимание на то, что в некоторых советских работах по автоматическому переводу термин «конфигурация» используется совсем в другом смысле.

2. Конфигурация не обязательно имеет только одну результирующую.

Примеры: *исключительно важный* $\Leftrightarrow$ *важный*, *ценный*, *полезный* ...; *не очень* $\Leftrightarrow$ *слишком*.

Обратим внимание на то, что выявление конфигураций в русском языке (а вероятно, и в других естественных языках) обязательно предполагает значительное огрубление наших представлений об этом языке, точнее, о грамматической правильности его фраз. А именно, признание указанных в примерах взаимозамещаемостей предполагает признание грамматической правильности, с одной стороны, цепочек типа *исключительно исключительно исключительно... важный* (поскольку в *исключительно важный* слово *важный* можно заменить на *исключительно важный* и т. д. сколько угодно раз), а с другой стороны, цепочек типа *очень исключительно важный* (из *очень важный*). Ср. также *это уже не очень* из *это уже слишком*, *не не очень влюбчив* из *не слишком влюбчив* и т. п.; аналогично *печальные забытые напевы* из *печальные напевы*, тогда как правильнее было бы *забытые печальные напевы* (см. выше, стр. 83—84). Подобная ситуация уже возникала раньше: строя примеры формальных грамматик, мы также были вынуждены признавать допустимость сколь угодно длинных цепочек типа *сладкая сладкая сладкая.. нежность* (стр. 57) или таких цепочек, как *сладкая горькая нежность*, *сладкая соль* и т. п.

б) **Индукционный шаг.** Допустим, что мы уже определили конфигурации всех рангов до ранга n включительно. Определим тогда конфигурацию ранга $n+1$. Это цепочка x, состоящая не менее чем из двух символов и удовлетворяющая следующим двум условиям:

1°. В словаре языка L имеется символ a, замещаемый на x.

2°. Цепочка x также замещаема на a, но не абсолютно, а с некоторыми ограничениями. Именно, в цепочке pxq ($pxq \in L$; p и q — произвольные цепочки) x можно заменить на a (т. е. $paq \in L$), если в цепочке pxq нет конфигураций ранга n или ниже,

которые пересекались бы с x, не содержась при этом в нем целиком. (Смысл этого условия состоит в том, что для возможности замены цепочки x на a цепочка x не должна иметь слева или справа от себя поясняющих слов, которые оставались бы «без хозяина» в случае замены x на a, см. примеры ниже.)

Символ a называется результирующей конфигурации x *).

Приведенное определение содержательно означает следующее. Во-первых, любую конфигурацию можно получить путем развертывания некоторого символа, а именно — ее результирующей. Во-вторых, любую конфигурацию можно «свернуть» к ее результирующей; однако конфигурацию ранга n можно «свертывать» только после того, как во фразе «свернуты» все конфигурации меньших рангов (от 1 до $n - 1$), имеющие с ней общую часть. Например, конфигурацию 2-го ранга *важный фактор* во фразе *...учитывает этот исключительно важный фактор* нельзя «свернуть», пока не свернута пересекающаяся с ней конфигурация 1-го ранга *исключительно важный*, так как иначе получилось бы неправильное ...**учитывает этот исключительно фактор*. Далее, конфигурацию 3-го ранга *воду пить* (с результирующей *пить*) во фразе *Грязную воду пить не следует* нельзя «свернуть», пока не «свернута» конфигурация 2-го ранга *грязную воду* (эллиптическую фразу *Грязную пить не следует*, которая получилась бы в противном случае, мы не считаем правильной).



Важно подчеркнуть, что данное определение конфигурации не является конструктивным — в нем не содержится никаких указаний на способ обнаруживать конфигурации в цепочках данного языка.

---

*) Следует иметь в виду, что в работах по математической лингвистике встречаются и другие варианты определения понятия конфигурации, не эквивалентные только что сформулированному (Кулагина 1958, Novotný 1965). Однако здесь имеются в виду именно варианты одного и того же понятия, а не совсем другое понятие, упомянутое в сноске к стр. 166.

Более того, неконструктивно и определение замещаемости: в самом деле, если бы мы захотели, пользуясь непосредственно определением, узнать, замещаема ли конкретная цепочка x на какую-то другую цепочку y (в языке L), нам пришлось бы перебрать все цепочки языка L, содержащие x, а таких цепочек в общем случае бесконечно много.

Оказывается, что можно утверждать и больше: существуют КС-языки, для которых нет алгоритма $\mathbf{T}_{1.2.11}$ распознавания конфигураций, т. е. алгоритма, способного для любой данной цепочки решить, является ли она конфигурацией (ср. выше, стр. 116).

С понятием конфигурации естественно связывается ряд других понятий — таких, как п р о с т а я к о н ф и г у р а ц и я (конфигурация, не содержащая в себе других конфигураций того же ранга, а тем самым — и конфигураций меньших рангов *); аналог минимальной неэлементарной составляющей) и н е п р и в о д и м а я ц е п о ч к а (цепочка, принадлежащая языку, т. е. являющаяся грамматически правильным предложением, и не содержащая в себе конфигураций, т. е. не допускающая свертываний. Неприводимыми цепочками являются, например, в русском языке безличные предложения типа *Морозит* или *Смеркалось*, а в английском и французском языках — любые предикативные пары типа *He reads* или *Il pleut*). На основе этих понятий строится определение к о н ф и г у р а ц и о н н о й х а р а к т е р и с т и к и языка. Конфигурационная характеристика есть совокупность двух перечней: это 1) перечень простых конфигураций, где для каждой конфигурации указана ее результирующая (или результирующие), и 2) перечень неприводимых цепочек. Вообще говоря, оба эти перечня могут быть бесконечными.

Любопытно, что конфигурационная характеристика полностью определяет язык в том смысле, что не может быть двух разных языков с одной и той

*) Конфигурация ранга n есть в то же время и конфигурация любого ранга, большего n; это непосредственно следует из определения конфигурации.

$T_{3.1}$ же конфигурационной характеристикой. Дело в том, что конфигурационная характеристика по существу есть не что иное, как набор элементарных цепочек и правил их комбинирования в цепочки языка. Ясно, что если наборы элементарных цепочек и правила их комбинирования одинаковы, то и полученные языки должны быть одинаковыми.

В случае, если перечни простых конфигураций и неприводимых цепочек конечны, соответствующий язык называется *конечно-характеризуемым*. (Интуитивно очевидно, что естественные языки являются конечно-характеризуемыми.) Выше (стр. 94) уже отмечалось, что все конечно-характеризуемые языки являются КС-языками. $T_{3.2}$

Мы не будем излагать здесь теорию конфигураций более подробно, так как нашей единственной целью было привести пример формального понятия (в нашем случае — конфигурации), выработанного математической лингвистикой в качестве уточняющей модели для одного из достаточно расплывчатых общелингвистических понятий (в нашем случае — для составляющей, точнее, для «типа составляющей», а не для конкретной составляющей в конкретной фразе).

Что касается формализации процедур лингвистического исследования, то эта область изобилует работами самых разнообразных направлений. Формализация процедур исследования находилась в центре внимания целой лингвистической школы — американской дескриптивной лингвистики — более тридцати лет (см., например, Harris 1963). Кроме того, сюда относятся работы по дешифровке текстов, написанных на неизвестных языках или неизвестной письменностью. (Деятельность дескриптивистов или специалистов-дешифровщиков по большей части не относится непосредственно к математической лингвистике, однако подготавливает почву для развития того ее раздела, о котором сейчас идет речь.) В настоящее время активно разрабатывается теория так называемой автоматической клас-

сификации, где ставится следующая задача: дано множество элементов текста, которым приписаны некоторые свойства; требуется автоматически разбить это множество на такие классы, чтобы элементы одного класса были с точки зрения указанных свойств более похожи друг на друга, чем элементы разных классов; подобные классы называют «сгустками» (clumps). Имеются также и работы других типов.

Мы рассмотрим здесь для примера предложенный в работе Сухотин 1962 алгоритм, который «умеет» в произвольном тексте на неизвестном (для алгоритма) языке обнаруживать гласные и согласные буквы (если, разумеется, текст написан буквенным письмом). При этом не используются никакие сведения о том конкретном языке, на котором написан текст. Единственная привлекаемая информация состоит в предположении, что общим свойством естественных языков является сравнительно равномерное распределение гласных и согласных в речи. Другими словами, это означает, что каков бы ни был язык, в тексте на нем чаще встречаются рядом (оказываются непосредственными соседями) гласная с согласной, а не гласная с гласной или согласная с согласной. (Например, в предыдущей фразе — за вычетом словосочетания в скобках — имеется 94 сочетания гласных с согласными, 31 сочетание согласных с согласными и 7 сочетаний гласных с гласными. Подсчет производится так, как если бы фраза была напечатана без пробелов; буквы *е*, *ю* и *я* считались гласными, а *й* — согласной.) Сформулированное допущение и лежит в основе алгоритма выявления гласных и согласных. Этот алгоритм разбивает все встречающиеся в тексте буквы на два класса Г и С таким образом, чтобы суммарная частота сочетаний типа ГС и СГ была больше, чем суммарная частота сочетаний ГГ и СС. Подобных разбиений может быть много, и среди них надо уметь выбрать наилучшее. Наилучшим же разбиением естественно считать такое, при котором разность суммарных частот сочетаний ГС и СГ, с одной

стороны, и ГГ и СС, с другой стороны, будет наибольшей, т. е. при котором классы Г и С будут максимально полярны.

Теперь мы сформулируем то же самое более точно. Пусть в тексте встречается n разных букв: a_1, $a_2, \ldots, a_n$. Допустим, что все эти буквы уже каким-нибудь способом разбиты на два класса — X и Y.

Для каждой пары букв $a_i\ a_j$ $(i, j = 1, 2, \ldots, n)$ подсчитаем число появлений этой пары в тексте (порядок a_i и a_j безразличен, т. е. $a_i\ a_j$ отождествляется с $a_j\ a_i$). Это число мы обозначим $F(a_i\ a_j)$; например, во фразе, для которой выше производились подсчеты («Другими словами,...»), F (*гл*) $=$ $= 6$, F (*ов*) $= 2$, F (*ру*) $= 1$, F (*щл*) $= 0$. Образуем сумму Σ_1 всех чисел $F(a_i\ a_j)$ для таких пар $a_i\ a_j$, где a_i и a_j принадлежат к разным классам ($a_i \in X$, $a_j \in Y$ или $a_i \in Y$, $a_j \in X$). Далее образуем аналогичную сумму Σ_2 всех чисел $F(a_i\ a_j)$ для таких пар $a_i\ a_j$, где a_i и a_j принадлежат одному классу (a_i, $a_j \in X$ или a_i, $a_j \in Y$). Разность $\Sigma_1 - \Sigma_2$ называется **функцией выгодности** и обозначается $P(X, Y)$. Наилучшим разбиением множества букв на классы X и Y будет то, при котором $P(X, Y)$ принимает максимальное значение. Это и есть определение наилучшего разбиения.

Поскольку множество букв конечно, существует тривиальный алгоритм нахождения максимума функции $P(X, Y)$ — полный перебор: надо произвести все возможные разбиения множества букв на классы X и Y, для каждого разбиения вычислить $P(X, Y)$ и среди полученных значений P (их конечное число!) выбрать наибольшее. Соответствующее разбиение и будет искомым *).

---

*) В общем случае функция $P(X, Y)$ может принимать одно и то же максимальное значение для нескольких разных разбиений.

Содержательно это означает наличие нескольких равноправных классификаций букв. Подобная ситуация, действительно, встречается в реальных языках: имеются буквы (на самом деле — звуки), которые по их сочетательным свойствам можно считать как гласными, так и согласными (так называемые полугласные, в некоторых языках—сонорные и т. п.).

Однако подобный алгоритм практически неосуществим: при алфавите, например, из 30 букв пришлось бы перебирать 2^{30} (т. е. больше миллиарда) различных разбиений. Поэтому были предложены (в работах Сухотин 1962, 1963) другие, более экономные алгоритмы, которые легко реализуются на электронной вычислительной машине. Эти алгоритмы были запрограммированы и опробованы в ряде экспериментов для русского, английского, немецкого, французского и испанского языков. Результаты экспериментов оказались вполне удовлетворительными; немногочисленные допущенные ошибки объясняются в основном условностями орфографии, например, тем, что один звук обозначается сочетанием букв (ср. нем. *sch* = [ʃ] и т. п.). Сами алгоритмы здесь не описываются, поскольку их конкретное устройство менее существенно для наших целей, чем постановка и принципы решения задачи.



У читателя может возникнуть вопрос: как соотносится сделанное выше (стр. 20—22) утверждение о неколичественном характере математической лингвистики и включение в нее работ (типа только что рассмотренной), где существенно используются количественные данные (в нашем случае — частоты). На самом деле, здесь нет противоречия. Во всех этих работах конечной целью является получение чисто качественных результатов — классификаций и т. п., а не количественных характеристик или закономерностей. Вычисления здесь представляют собой промежуточный этап, т. е. при всей своей важности остаются вспомогательным средством.

Использование разнообразных функций выгодности кладется в основу также и других алгоритмов выделения и классификации языковых единиц. Например, в статье Сухотин 1963 изложен алгоритм членения написанного без пробелов текста на морфы. Функция выгодности строится здесь исходя из того соображения, что суммарная частота совместной встречаемости букв, входящих в одну морфу, больше, чем суммарная частота совместной

встречаемости букв из разных морф. Ср. также алгоритм обнаружения синтаксически связанных слов в тексте на неизвестном языке (Сухотин 1963).

Здесь наше изложение подходит к концу. Нам остается попрощаться с читателем, взявшим на себя труд дойти вместе с нами до этих строк. Впрочем, если его терпение еще не истощилось, он может обратиться к Заключению и к Приложениям I и II: в первом из них дана сводка важнейших математических результатов теории грамматик, а второе представляет собой комментированный словарик терминов математической лингвистики. Благодарим за внимание!

Заключение

В самом начале книги, на стр. 16, было сказано: Математическая лингвистика есть математическая дисциплина, «обращенная» в сторону естественных языков и лингвистики. Теперь, когда наше изложение закончено, мы можем и даже, по-видимому, должны уточнить эту недостаточно конкретную формулировку, для чего мы рассмотрим вопрос о соотношении между математической лингвистикой и «просто» лингвистикой, с одной стороны, и между математической лингвистикой и математикой — с другой.

Лингвистика, или, точнее, теоретическая лингвистика (в противоположность описательной), может быть охарактеризована как наука, занимающаяся построением формальных моделей естественных языков и методологией построения таких моделей. Она строит м о д е л и, поскольку теоретическое изучение любого объекта (в данном случае — языка) не может быть ничем иным, как созданием его моделей, а эти модели должны быть ф о р м а л ь н ы м и — при изучении естественных языков рассматривать какие-либо иные «модели» нецелесообразно. В то же время читатель, видимо, заметил, что и ма-

тематическая лингвистика занимается разработкой формальных конструкций, призванных служить моделями естественных языков. Однако здесь имеется весьма существенное различие, заключающееся в следующем. Лингвистика разрабатывает теорию языка как такового, т. е. общую модель языка, а также строит конкретные модели конкретных языков или их фрагментов; математическая лингвистика изучает на абстрактном уровне самые общие свойства таких моделей, исследуя средства и методы их построения. Другими словами, математическая лингвистика создает и исследует абстрактные (можно было бы сказать «умозрительные» или, пользуясь устаревшим термином, «спекулятивные») модели конкретных лингвистических моделей, т. е. «модели второго уровня». Математическая лингвистика занимается, так сказать, материалами, инструментами, общей теорией строительства; лингвистика же непосредственно возводит здания. Если лингвистика есть теория конкретных языков и языка вообще, то математическая лингвистика — это теория строения этой теории, т. е. метатеория, или металингвистика. Строя модели естественных языков, лингвистика неминуемо пользуется некоторым определенным языком (который является метаязыком по отношению к описываемым естественным языкам). Разработкой и чисто формальным исследованием этого языка лингвистики и занимается математическая лингвистика.

Разумеется, абсолютно четкую границу между лингвистикой и математической лингвистикой провести невозможно. Лингвист, разрабатывающий конкретную модель языка, может заняться изучением в общем виде тех средств, которыми он пользуется; специалист по математической лингвистике, исследуя свои чисто абстрактные конструкции, также может заинтересоваться их возможной интерпретацией, т. е. их приложением к конкретным языкам. В результате во многих реальных работах сферы лингвистики и математической лингвистики настолько тесно переплетаются, что отнести работу к одной из этих областей затруднительно. Тем не менее

сформулированное выше противопоставление в принципе сохраняет силу: лингвистика сосредоточена в первую очередь на описании общих законов и свойств языка, а также отдельных конкретных языков, математическая лингвистика — на изучении средств и способов этого описания.

Что же касается взаимоотношений между математической лингвистикой и математикой, то, как должно быть видно из предшествующего изложения, первая есть часть второй. При этом из всех математических дисциплин математическая лингвистика стоит ближе всего к математической логике и значительной своей частью (здесь имеется в виду теория грамматик) просто входит в нее. А поскольку математическая логика занимается изучением языка математики, то и математическая лингвистика применима и применяется для этой цели. В частности, любопытно отметить, что категориальные грамматики были введены К. Айдукевичем еще в 1935 г. для исследования строения некоторых математических систем. В последнее время формальные грамматики широко используются для задания (описания) так называемых языков программирования (искусственных языков, применяемых для записи алгоритмов с целью придать им стандартную форму, которая облегчала бы их ввод в вычислительную машину).

Приложение I

Сводка математических результатов

Ниже формулируется ряд наиболее существенных, по мнению авторов, утверждений, относящихся к различным разделам математической лингвистики. Большинство из них (но не все) содержится в тексте книги (ср. замечание на стр. 13). Для большей обозримости здесь повторяются также некоторые определения. После каждой теоремы указаны соответствующие страницы книги, а также работы, содержащие доказательство теоремы.

1. *Порождающие грамматики*

Основные определения. Порождающая грамматика (ПГ) есть упорядоченная четверка $\langle \boldsymbol{V}, \boldsymbol{V}_1, \boldsymbol{I}, \boldsymbol{S} \rangle$, где $\boldsymbol{V}$ и $\boldsymbol{V}_1$ — непересекающиеся конечные множества символов ($\boldsymbol{V}$ — основной словарь, $\boldsymbol{V}_1$ — вспомогательный словарь), $\boldsymbol{I}$ — элемент $\boldsymbol{V}_1$ (начальный символ), а $\boldsymbol{S}$ — конечное множество правил вида $A \to B$, где A и B — произвольные цепочки над словарем $\boldsymbol{V} \cup \boldsymbol{V}_1$ (схема грамматики). Определения непосредственной выводимости, выводимости и вывода см. стр. 43; определение языка, порождаемого грамматикой, см. стр. 46.

1.1. Соотношение между различными классами порождающих грамматик.

Неукорачивающая грамматика — ПГ, в каждом правиле которой правая часть не короче левой: $l(A) \leqslant l(B)$.

Грамматика непосредственно составляющих (НС-грамматика) — ПГ, каждое правило которой имеет вид $Z_1CZ_2 \to Z_1WZ_2$, где C — вспомогательный символ, а W — непустая цепочка.

Контекстно-свободная грамматика (КС-грамматика) — ПГ, каждое правило которой имеет вид $C \to W$, где C — вспомогательный символ, а W — непустая цепочка.

Бинарная КС-грамматика — КС-грамматика, в каждом правиле которой правая часть содержит не более двух символов.

Автоматная грамматика (А-грамматика) — КС-грамматика, в каждом правиле которой правая часть либо состоит из одного символа, либо имеет вид bB, где b — основной символ, а B — вспомогательный символ.

КС-грамматика с ограниченной памятью — КС-грамматика, для которой существует такое число K, что какова бы ни была терминальная цепочка x, выводимая в этой грамматике, найдется такой вывод этой цепочки из начального символа в данной грамматике, что любая промежуточная цепочка этого вывода содержит вспомогательные символы не более чем на K последних местах.

КС-грамматика с независимым двусторонним развертыванием — см. определение на стр. 102—103.

Теорема 1.1.1. Класс языков, порождаемых грамматиками, совпадает с классом рекурсивно перечислимых языков (стр. 49; Davies 1958 и Гладкий 1966).

Теорема 1.1.2. Класс языков, порождаемых неукорачивающими грамматиками, совпадает с классом НС-языков; более того, для всякой неукорачивающей грамматики можно эффективно построить эквивалентную ей НС-грамматику (стр. 54; Хомский 1966, стр. 165, и Гладкий 1966, стр. 72).

Теорема 1.1.3. Существуют НС-языки, не являющиеся КС-языками, например, $\{a^n b^n a^n\}$ (стр. 65; Bar-Hillel — Perles — Shamir 1961 и Гладкий 1966, стр. 91).

Теорема 1.1.4. Для всякой КС-грамматики можно эффективно построить эквивалентную ей бинарную КС-грамматику (стр. 67; Гладкий 1966, стр. 83).

Теорема 1.1.5. Существуют КС-языки, не являющиеся А-языками, например, $\{a^n b^n\}$ (стр. 71; Гладкий 1966, стр. 102).

Теорема 1.1.6. Для всякой КС-грамматики с ограниченной памятью можно эффективно построить эквивалентную ей А-грамматику (стр. 100).

Теорема 1.1.7. Для всякой КС-грамматики с независимым двусторонним развертыванием можно эффективно построить эквивалентную ей А-грамматику (стр. 104).

Теорема 1.1.8. а) Класс НС-языков замкнут относительно операций теоретико-множественного объединения и пересечения *). б) Класс КС-языков замкнут относительно операции теоретико-множественного объединения и не замкнут относительно операций теоретико-множественного пересечения и дополнения. в) Класс А-языков замкнут относительно операций теоретико-множественного объединения, пересечения и дополнения (см. Гладкий 1966, стр. 104, и Хомский 1966).

1.2. Алгоритмические проблемы теории порождающих грамматик.

Теорема 1.2.1. В классе всех грамматик ни одно нетривиальное свойство языков, порождаемых грамматиками, не распознаваемо (см. стр. 113; Гладкий 1966, стр. 161) **).

Теорема 1.2.2. В классе НС-грамматик свойство порождать язык, содержащий данную цепочку, распознаваемо (стр. 114).

---

*) Вопрос о замкнутости класса НС-языков относительно операции дополнения остается открытым; имеется в виду дополнение до множества всех непустых цепочек над данным терминальным словарем.

**) Эта теорема есть простое следствие теоремы Райса, доказательство которой имеется, например, в книге А. И. Мальцева «Алгоритмы и рекурсивные функции», М., 1965, стр. 147.

Теорема 1.2.3. В классе НС-грамматик следующие свойства нераспознаваемы: порождать данный (произвольный) НС-язык, порождать конечный язык, порождать КС-язык, порождать язык с замещаемостью x на y, где x и y — произвольные (фиксированные) цепочки (стр. 114; Гладкий 1966, стр. 163—171; Гладкий 1964б).

Теорема 1.2.4. В классе КС-грамматик следующие свойства распознаваемы: порождать пустой язык, порождать конечный язык, порождать хотя бы одну цепочку, содержащую вхождение данной цепочки x (стр. 115; Bar-Hillel — Perles — Shamir 1961; Гладкий 1966, стр. 85—89).

Теорема 1.2.5. В классе КС-грамматик следующие свойства нераспознаваемы: порождать А-язык, порождать полный язык (язык, содержащий все терминальные цепочки над данным словарем), иметь эквивалентную КС-грамматику, приписывающую каждой терминальной цепочке только одну синтаксическую структуру (стр. 115; Bar-Hillel — Perles — Shamir 1961; Гладкий 1966, стр. 171—181).

Теорема 1.2.6. В классе А-грамматик следующие свойства распознаваемы: порождать полный язык, порождать язык с конечным дополнением (стр. 115 Bar-Hillel — Perles — Shamir 1961).

Теорема 1.2.7. Не существует алгоритма, позволяющего по любой паре КС-грамматик узнать, являются ли они эквивалентными (стр. 115; Bar-Hillel — Perles — Shamir 1961, Гладкий 1966).

Теорема 1.2.8. Существует алгоритм, позволяющий по любой паре А-грамматик узнать, являются ли они эквивалентными (стр. 115; Bar-Hillel — Perles — Shamir 1960, Гладкий 1964а).

Теорема 1.2.9. Существует КС-грамматика Г, для которой невозможен алгоритм, позволяющий по любой паре цепочек x, y узнать, замещаемо *) ли x на y в языке L (Г) (стр. 116; Гладкий 1965).

Теорема 1.2.10. Для всякой А-грамматики Г существует алгоритм, позволяющий по любой паре

*) Определения замещаемости и конфигурации см. стр. 166—168.

цепочек x, y узнать, замещаемо ли x на y в языке L (Г) (стр. 116; Гладкий 1964а).

Теорема 1.2.11. Существует КС-грамматика Г, для которой невозможен алгоритм, позволяющий по любой цепочке узнать, является ли она конфигурацией*) языка L (Г) (стр. 169; Лучкин 1966).

Теорема 1.2.12. Для всякой А-грамматики Г существует алгоритм, позволяющий по произвольной цепочке x, символу a и числу n узнать, является ли x конфигурацией ранга n языка L(Г) с результирующей a (Гладкий 1964а).

Теорема 1.2.13. Для всякой неукорачивающей грамматики существует алгоритм, позволяющий по любой цепочке узнать, выводима ли она в данной грамматике из ее начального символа (стр. 50).

1.3. Оценки сложности вывода.

Теорема 1.3.1. Для всякой неукорачивающей грамматики Г имеет место неравенство $\tau_\Gamma(n) < p^{n+1}$, где τ_Γ — временна́я сигнализирующая функция грамматики Г (см. стр. 116), а p — общее число основных и вспомогательных символов этой грамматики (стр. 117).

Теорема 1.3.2. Для всякой КС-грамматики Г имеет место неравенство $\tau_\Gamma(n) \leqslant 2\,Cn$, где C — число вспомогательных символов грамматики Г (стр. 117).

Теорема 1.3.3. Язык $\{xqx'\}$ (см. стр. 91) не может быть порожден никакой НС-грамматикой, у которой временна́я сигнализирующая функция $\tau_\Gamma(n)$ по порядку меньше чем n^2 (стр. 118; Гладкий 1964в.).

Теорема 1.3.4. Класс языков, порождаемых грамматиками с ограниченным растяжением, совпадает с классом НС-языков; более того, для всякой грамматики с ограниченным растяжением можно эффективно построить эквивалентную ей НС-грамматику (стр. 105; Гладкий 1964 в).

---

*) В статье Гладкий 1963б вместо термина «грамматики с ограниченным растяжением» используется термин «грамматики с линейной памятью».

2. *Другие типы формальных грамматик*

Определение категориальной грамматики (К-грамматики) см. стр. 125; определение автомата с магазинной памятью (М-автомата) см. стр. 148—149.

Теорема 2.1. Всякий К-язык есть КС-язык; более того, для всякой К-грамматики можно эффективно построить эквивалентную ей КС-грамматику (см. стр. 133; Bar-Hillel — Gaifman — Shamir 1960; Гладкий 1966, стр. 123).

Теорема 2.2. (Гайфман). Всякий КС-язык есть К-язык; более того, для всякой КС-грамматики можно эффективно построить эквивалентную ей К-грамматику (см. стр. 133; Bar-Hillel — Gaifman — Shamir 1960; Гладкий 1966, стр. 123).

Теорема 2.3. (Хомский). Всякий М-язык есть КС-язык; более того, для всякого М-автомата можно эффективно построить эквивалентную ему КС-грамматику (стр. 149; Хомский 1966).

Теорема 2.4. Всякий КС-язык есть М-язык; более того, для всякой КС-грамматики можно эффективно построить эквивалентный ей М-автомат (стр. 149; Хомский 1966).

3. *Конфигурации*

Цепочка x замещаема на цепочку y в языке L [обозначение: $x \Rightarrow y$ (L)], если для любых двух цепочек p и q из $pxq \in L$ следует $pyq \in L$. Цепочка x взаимозамещаема с цепочкой y в языке L, если $x \Rightarrow y$ (L) и $y \Rightarrow x$ (L) [обозначение: $x \Leftrightarrow y$ (L)].

Если x — цепочка, состоящая не менее чем из двух символов и a — символ, то x называется конфигурацией 1-го ранга (языка L) с результирующей a, если $x \Leftrightarrow a$ (L). Если определены конфигурации рангов $\leqslant n$ и если x — цепочка, состоящая не менее чем из двух символов, и a— символ, то x называется конфигурацией ранга $n+1$ (языка L) с результирующей a, если: 1) $a \Rightarrow x$ (L) и 2) для любых двух цепочек p и q из того, что pxq принадлежит L и не содержит вхождений конфигураций рангов $\leqslant n$,

пересекающихся с выделенным вхождением x, но не содержащихся в нем целиком, следует, что paq принадлежит L (см. стр. 166).

Простая конфигурация (языка L) — конфигурация, не содержащая вхождений других конфигураций того же ранга; неприводимая цепочка языка L — цепочка, принадлежащая языку L и не содержащая вхождений конфигураций языка L.

Конфигурационная характеристика языка L есть упорядоченная пара $\langle Б\ (L),\ П\ (L)\rangle$, где $Б\ (L)$ — множество всех неприводимых цепочек языка L и $П\ (L)$ — множество всевозможных упорядоченных пар вида $\langle a,\ x\rangle$, где x — простая конфигурация языка L и a — ее результирующая.

Если $Б\ (L)$ и $П\ (L)$ конечны, то L называется конечно характеризуемым языком.

Т е о р е м а 3.1. Если L_1 и L_2 — языки над одним и тем же словарем и $Б\ (L_1) = Б\ (L_2)$, $П\ (L_1) = П\ (L_2)$, то $L_1 = L_2$; иначе говоря, язык над данным словарем полностью определяется своей конфигурационной характеристикой (стр. 170; Гладкий 1963а).

Т е о р е м а 3.2. Класс конечно характеризуемых языков содержится в классе КС-языков; более того, по конечной конфигурационной характеристике языка можно эффективно построить КС-грамматику, порождающую этот язык (стр. 170; Гладкий 1963а).

Т е о р е м а 3.3. Существуют КС-языки (и даже А-языки), не являющиеся конечно характеризуемыми (Гладкий 1963а).

Т е о р е м а 3.4. Всякий А-язык без омонимии*) является конечно характеризуемым языком; обратное неверно (Гладкий 1963а).

Т е о р е м а 3.5. Существуют КС-языки, в которых для любого натурального числа n имеются конфигурации ранга n, не являющиеся конфигурациями ранга $n - 1$ (Лучкин 1966).

---

*) Язык, порождаемый такой А-грамматикой, что никакие два ее правила, имеющие разные левые части, не могут содержать в правых частях один и тот же основной символ.

Приложение II

Краткий словарь основных терминов математической лингвистики

Автомат с магазинной памятью (М-автомат) — push-down store automaton/machine. См. стр. 136.

Автоматная грамматика (А-грамматика) — finite state grammar. Калькирующий русский термин «грамматика с конечным числом состояний» неудачен, поскольку грамматик «с бесконечным числом состояний» не бывает. Английский термин на самом деле шире по значению, чем его русская калька: он означает нечто вроде «грамматика, имеющая определенное отношение к конечному числу состояний [чего-то]», причем имеется в виду finite state automaton, т. е. эквивалентный грамматике автомат с конечным числом состояний (конечный автомат). Именно этой эквивалентностью и объясняется предлагаемый термин; совсем точным названием было бы «конечно-автоматная грамматика». См. стр. 69 и сл.

Автоматный язык (А-язык) — finite state language. Калька «язык с конечным числом состояний» неприемлема, так как языки в смысле теории формальных грамматик вообще не могут иметь «состояний». Как и в предыдущем случае, английский термин означает, собственно говоря, «язык, имеющий определенное отношение к конечному числу состояний [чего-то]», т. е. конечного автомата. Точнее было бы говорить о конечно-автоматных языках, ср. выше. См. стр. 70.

Взаимозамещаемость — mutual substitutability/ /interchangeability. См. стр. 166.

Вспомогательный символ — non-terminal symbol. Употребляется также термин «нетерминальный символ». См. стр. 37.

Вспомогательный словарь — non-terminal vocabulary. Употребляется также термин «нетерминальный словарь». См. стр. 37.

Вывод — derivation. Употребляемый иногда термин «деривация» абсолютно неприемлем, так как

соответствующее понятие давно обозначается в русской математической литературе именно термином «вывод». Кроме того, за словом «деривация» в лингвистике уже закрепился другой смысл: аффиксальное словообразование. См. стр. 43.

Выводимость — derivability. См. стр. 43.

Главная категория — main category. См. стр. 125.

Грамматика — grammar. См. стр. 23.

Грамматика непосредственно составляющих (НС-грамматика) — phrase structure grammar. Русская калька «фразово-структурная грамматика» неприемлема, поскольку англ. phrase означает именно «составляющая» (или «словосочетание»), но ни в коем случае не «фраза». Кроме того, термины «непосредственно составляющие», «анализ по непосредственно составляющим» и т. п. давно приняты в лингвистике. Они возникли как буквальный перевод английского термина immediate constituent (analysis, approach, description). По-видимому, удобнее было бы говорить просто о грамматике составляющих, опуская определение «непосредственно», поскольку никаких других составляющих не бывает. Однако в связи с общеупотребительностью термина «непосредственно составляющие» мы предпочли его сохранить. См. стр. 54.

Грамматика с ограниченным замедлением — см. стр. 119.

Грамматика с ограниченным растяжением — см. стр. 105.

Дегуманизация — dehumanization. См. стр. 9.

Деривация — см. **Вывод.**

Замещаемость — substitutability. См. стр. 166.

Категориальная грамматика — categorial grammar. См. стр. 122.

Категория — category. См. стр. 124—125.

Конечно характеризуемый язык — см. стр. 170.

Контекстно-свободная грамматика (КС-грамматика) — context-free grammar/simple phrase structure grammar. Используется также термин «бесконтекстная грамматика» (представляющий собой бо-

лее удачный в стилистическом отношении перевод английского выражения). См. стр. 63.

Контекстно-свободное правило — context-free rule. См. стр. 63.

Контекстно-связанная грамматика — context-bound/context-dependent/context-sensitive grammar. См. стр. 63.

Контекстно-связанное правило —context-bound/context-dependent/context-sensitive rule. См. стр. 63.

Конфигурационная характеристика — см. стр. 169.

Конфигурация — см. стр. 166—168.

Магазин — push-down store. См. стр. 141.

Начальный символ — initial symbol. См. стр. 38.

Нетерминальный символ — см. **Вспомогательный символ.**

Нетерминальный словарь — см. **Вспомогательный словарь.**

НС-грамматика — см. Грамматика непосредственно составляющих.

Основной/терминальный символ — terminal symbol. См. стр. 36.

Основной/терминальный словарь — terminal vocabulary. См. стр. 35.

Правила подстановки — rewriting rules. Часто применяемый русский термин «правила переписывания», калькирующий англ. rewriting rules, совершенно неприемлем в силу омонимичности глагола «переписывать»: англ. *copy* «переписывать (без изменения)» и англ. *rewrite* «переписывать (заново, быть может, в ином виде)». Термин «переписывать» приходится использовать в теории грамматик в смысле англ. *copy* — когда имеется в виду перенос символа из одной цепочки в другую без изменения. Кроме того, термин «правила подстановки» давно используется в русской математической литературе именно в нужном смысле. См. стр. 38.

Предсказуемостный/предиктивный анализатор — predictive analyzer. См. стр. 136.

Приписывающая функция — assignment function. См. стр. 125.

Сигнализирующая функция временна́я — см. стр. 116.

Сигнализирующая функция емкостная — см. стр. 119.

Словарь — vocabulary. См. также **Основной словарь, Вспомогательный словарь.** Наряду с термином «словарь» иногда употребляется также термин «алфавит» (англ. alphabet).

Сокращение — cancellation. См. стр. 125.

Составляющая — constituent/phrase. См. стр. 164 и сл.

Цепочка — string. Употребление в данном смысле термина «фраза», встречающееся в некоторых работах, является неудачным, поскольку в русской лингвистической литературе фразами называют не любые цепочки символов, а только некоторые, имеющие определенное строение ($\approx$ предложения). См. стр. 23.

Циклическое правило — recursive rule. Термин «рекурсивное правило» представляется неудобным, поскольку за словом «рекурсивный» в математике закрепилось другое значение. См. стр. 57.

Эквивалентность грамматик — см. стр. 47.

Элементарная категория — primitive category. См. стр. 124.

Язык с конечным числом состояний — finite state language — см. **Автоматный язык.**

Язык, порождаемый грамматикой,— См. стр. 47.

Литература

А б а е в В. И. Лингвистический модернизм как дегуманизация науки о языке.— «Вопросы языкознания», 1965, № 3, 22—43.

А п р е с я н Ю. Д. Современные методы изучения значений и некоторые проблемы структурной лингвистики. В кн.: «Проблемы структурной лингвистики», М., 1963.

А п р е с я н Ю. Д. Идеи и методы современной структурной лингвистики (краткий очерк), М., 1966.

А р с е н т ь е в а Н. Г. О двух способах порождения предложений русского языка.— В сб.: «Проблемы кибернетики», вып. 14, 1965, 189—218.

Б а р - Х и л л е л И. Разрешающие процедуры для структуры естественных языков.— В кн.: «Математическая лингвистика», М., 1964, 108—121.

Б а р - Х и л л е л И. Некоторые новые результаты в теоретической лингвистике.— В кн.: «Математическая логика и ее применения», М., 1965, 273—280.

Г и л ь б е р т Д., А к к е р м а н В. Основы теоретической логики, М., 1947.

Г л а д к и й А. В. Конфигурационные характеристики языков.— В сб.: «Проблемы кибернетики», вып. 10, 1963 (а), 251—260.

Г л а д к и й А. В. Грамматики с линейной памятью.— «Алгебра и логика», 1963 (б), 2, № 5, 43—55.

Г л а д к и й А. В. Алгоритм распознавания конфигураций для класса автоматных языков.— В сб.: «Проблемы кибернетики», вып. 12, 1964 (а), 243—245.

Г л а д к и й А. В. Алгоритмическая природа инвариантных свойств грамматик непосредственно составляющих.— «Алгебра и логика», 1964 (б), 3, № 2, 17—31.

Г л а д к и й А. В. О сложности вывода в грамматиках непосредственно составляющих. — «Алгебра и логика», 1964(в), 3, 5—6, 29—44.

Г л а д к и й А. В. Некоторые алгоритмические проблемы для контекстно-свободных грамматик.— «Алгебра и логика», 1965, 4, № 3, 3—13.

Г л а д к и й А. В. Лекции по математической лингвистике для студентов НГУ, Новосибирск, Изд. НГУ, 1966.

Г л а д к и й А. В., Р ы б а к о в а М. В., Ш е д ь к о Т. И. Схема семантического языка-посредника для записи математических текстов, М., 1961 (ВИНИТИ АН СССР. Доклады на конференции, вып. 10).

Глисон Г. Введение в дескриптивную лингвистику, М., 1959.

Жолковский А. К., Леонтьева Н. Н., Мартемьянов Ю. С. О принципиальном использовании смысла при машинном переводе.— В кн.: «Машинный перевод», Труды ИТМ и ВТ, вып. 2, М., 1961, 17—46.

Жолковский А. К., Мельчук И. А. О семантическом синтезе.— В сб.: «Проблемы кибернетики», вып. 19, 1967, 177—238.

Зализняк А. А. Русское именное словоизменение, М., 1967.

Зализняк А. А. Синтаксические категории имени, 1969 (в печати).

Ингве В. Гипотеза глубины.— В кн.: «Новое в лингвистике», вып. IV, М., 1965, 126—138.

Кузнецов А. В., Падучева Е. В., Ермолаева Н. М. Об информационном языке для геометрии и алгоритме перевода с русского языка на информационный.—В кн.: «Лингвистические исследования по машинному переводу, вып. 2, М., 1961, 40—73. См. также сб. «Машинный перевод и прикладная лингвистика», вып. 5, 1961; 3—21; вып. 6, 1961, 9—18.

Кулагина О. С. Об одном способе определения грамматических понятий на базе теории множеств.—В сб.: «Проблемы кибернетики», вып. 1, 1958, 203—214.

Ламбек И. Математическое исследование структуры предложений.— В кн.: «Математическая лингвистика», М., 1964, 47—68.

Ломковская М. В. Исчисление, порождающее ядерные русские предложения.— «Научно-техническая информация», 1965, № 7, 35—41 (часть I); 1965, № 9, 37—40 (часть II); 1966, № 11, 56—65 (часть III).

Лучкин В. Д. О рангах конфигураций контекстно-свободных языков.— «Алгебра и логика», 1966, 5, № 3, 59—70.

Маркус С. Грамматический род и его логическая модель.— В кн.: «Математическая лингвистика», М., 1964.

Мастерман М. Изучение семантической структуры текста для машинного перевода с помощью языка-посредника.— В кн.: «Математическая лингвистика», М., 1964, 177—213.

Мельчук И. А. Автоматический анализ текстов (на материале русского языка).—В кн.: «Славянское языкознание» (Доклады советской делегации. V Международный съезд славистов), М., 1963, 477—509.

Мельчук И. А. Автоматический синтаксический анализ, I, Новосибирск, 1964.

Мельчук И. А. Порядок слов при автоматическом синтезе русского текста (Предварительное сообщение).— «Научно-техническая информация», 1965, № 12, 36—41.

Падучева Е. В. Некоторые вопросы перевода с информационно-логического языка на русский.— «Научно-техническая информация», 1964(а), № 2, 20—27.

Падучева Е. В. О способах представления синтаксической структуры предложения.— «Вопросы языкознания», 1964(б), № 2, 99—113.

Падучева Е. В. Опыт логического анализа значения союза *или*.— «Философские науки. НДВШ», 1964(в), № 6, 145—148.

Падучева Е. В. Синтез сложных предложений с однозначной синтаксической структурой (при переводе с информационно-логического языка на русский).— «Научно-техническая информация», 1964(г), № 6, 43—49.

Падучева Е. В. О связях глубины по Ингве со структурой дерева подчинений.— «Научно-техническая информация», 1967, № 6, 38—43.

Паркер-Роудс А. Ф. О последних работах в области машинного перевода методом тезауруса и языка-посредника.— В кн.: «Математическая лингвистика», М., 1964, 229—245.

Ревзин И. И. Метод моделирования и типология славянских языков, М., 1967.

Сухотин Б. В. Экспериментальное выделение классов букв с помощью электронной вычислительной машины.— В кн.: «Проблемы структурной лингвистики», М., 1962, 198—206.

Сухотин Б. В. Алгоритмы лингвистической дешифровки.— В кн.: «Проблемы структурной лингвистики», М., 1963, 75—101.

Тарский А. Введение в логику и методологию дедуктивных наук, М., 1948.

Трахтенброт Б. А. Алгоритмы и машинное решение задач, М., 1960.

Успенский В. А. Одна модель для понятия фонемы.— «Вопросы языкознания», 1964, № 6, 39—53.

Хомский Н. Синтаксические структуры.—В кн.: «Новое в лингвистике», вып. II, 1962, 412—527.

Хомский Н. Логические основы лингвистической теории.— В кн.: «Новое в лингвистике», вып. IV, 1965(а), 465—575.

Хомский Н. О понятии «правило грамматики».—В кн.: «Новое в лингвистике», вып. IV, 1965(б), 34—65.

Хомский Н. Формальные свойства грамматик.—В кн.: «Кибернетический сборник». Новая серия, вып. 2, 1966, 121—230.

Хомский Н., Миллер Дж. Введение в формальный анализ естественных языков.—В кн.: «Кибернетический сборник». Новая серия, вып. 1, 1965, 229—290.

Хэррис З. С. Совместная встречаемость и трансформация в языковой структуре.—В кн.: «Новое в лингвистике», вып. II, 1962, 528—636.

Шрейдер Ю. А. Характеристики сложности структуры текста.— «Научно-техническая информация», 1966, № 7, 34—41.
Bar-Hillel Y.*) A quasi-arithmetical notation for syntactic description.— «Language», 1953, 29, № 1, 47—58.
Bar-Hillel Y. Some linguistic obstacles to machine translation.— In. «Advances in computers», vol. 1, N. Y.— London, 1960, 146—157.
Bar-Hillel Y., Gaifman C., Shamir E. On categorial and phrase-structure grammars.— «Bulletin of the Research Council of Israel», 1960, 9F, 1—16.
Bar-Hillel Y., Shamir E. Finite state languages: formal representation and adequacy problems.— «Bulletin of the Research Council of Israel», 1960, 8F, № 3, 155—166.
Bar-Hillel Y., Perles M., Shamir E. On formal properties of simple phrase structure grammars.— «Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung», 1961, 14, № 2, 143—172.
Bobrow D. G. Syntactic analysis of English by computer — a survey.— In. «AFIPS conference proceedings», 24, Baltimore — London, 1963, 365—387.
Charney E. K. On the semantic interpretation of linguistic entities that function structurally. Teddington (Nat. physical lab. Paper 8), 1961.
Charney E. K. On the probleme of sentence synonymy.— «Quarterly progress report», № 66 (Mass. Inst. of technology), Cambridge (Mass.), 1962, 289—293.
Charney E. K. Structural semantic foundations for a theory of meaning, Chicago, 1966.
Davies M. Computability and unsolvability, N. Y., 1958.
Döhmann K. Zur Semantik und Etymologie der sprachlichen Darstellung der dyadischen Funktoren.— «Studium generale», 1966, 19, № 7, 398—401.
Freudenthal H. LINCOS. Design of a language for cosmic intercourse, Part 1, Amsterdam, 1960.
Harris Z. S. Structural linguistics, Chicago — London, 1963.
Hays D. Grouping and dependency theories.—In. «Proceedings of the National Symposium on Machine Translation», Englewood Cliffs (N. J.), 1961, 258—266.
Hays D. G. Automatic language data processing.—In.«Computer applications in behavioral sciences», Englewood Cliffs (N. J.), 1962, 394—421.
Hays D. Dependency theory: a formalisation and some observations.— «Language», 1964, 64, № 4.

*) Указанные здесь работы Бар-Хиллела и его соавторов перепечатаны в книге Y. Bar-Hillel «Language and information», Reading (Mass.) etc., 1964.

Kuno S. The multiple-path syntactic analyzer for English.— «Mathematical linguistics and automatic translation» (Computation lab. Harvard univ.), Report № NSF-9, vol. 1, Cambridge (Mass.), 1963.

Kuno S., Oettinger A. Syntactic structure and ambiguity of English.— In: «AFIPS proceedings of the Fall joint computer conference...», 24, Baltimore — London, 1963, 397—418.

Marcus S. Le genre grammatical et son modèle logique.— «Cahiers de linguistique théorique et appliquée», 1962, 1, 103—122.

Marcus S. Modèles mathématiques pour la catégorie grammaticale du cas.— «Revue de mathématiques pures et appliquées, 1963, 3, № 4, 585—610.

Mendelson E. Introduction to mathematical logic, Princeton — Toronto — New York — London, 1964.

Novotný M. Über endlich charakterisierbaren Sprachen.— «Spisy prirodověd. fak. univ. v Brně», 1965, № 10, 495—502.

Plath W. J. Mathematical linguistics.— In: «Trends in European and American linguistics, 1930—1960...», Utrecht—Antwerpen, 1961, 21—57.

Plath W. J. Multiple-path syntactic analysis of Russian.— «Mathematical linguistics and automatic translation» (Computation lab. Harvard univ.), Report № NSF-12, Cambridge (Mass.), 1963.

Postal P. M. Limitations of phrase structure grammars.— In: «The structure of language. Readings in the philosophy of language», Englewood Cliffs (N. J.), 1964, 137—151.

Reichenbach H. Elements of symbolic logic, N. Y., 1960.

Russel B. An inquiry into meaning and truth, N. Y., 1940.

Quine W. V. Word and object, New York — London, 1960.

Quine W. V. Logic as a source of syntactical insights.— In: «Structure of language and its mathematical aspects», Providence (Rhode Island), 1961, 1—5.

Tesnière L. Eléments de syntaxe structurale, P., 1959.

Tosh L. W. Syntactic translation, The Hague, 1965.

Varga D. Yngve's hypothesis and some problems of the mechanical analysis.— «Computational Linguistics», 1964, III, 47—74.

Weinreich U. On the semantic structure of language.— In: «Universals of language», Ed. J. H. Greenberg, Cambridge (Mass.), 1963, 114—171.

Yngve V. H. A model and a hypothesis for language structure.— «Proceedings of American phylosophical society», 1960, 104, № 5, 444—466.

Yngve V. H. Random generation of English sentences, Teddington (National physical laboratory. Paper 6), 1961.

Опечатки

Стр.	Строка	Напечатано	Следует читать
28	6 сн.	«ся»	«$\overline{\text{ся}}$»
28	5 сн.	«ся-ся»	«ся-$\overline{\text{ся}}$»
125	4 св.	$[\Phi\diagdown\Psi]$ и	$[\Phi/\Psi]$ и
125	10 сн.	$[\Phi\,[\Phi\diagdown\Psi]$	$\Phi\,[\Phi\diagdown\Psi]$
125	8 сн.	$\Phi\diagdown\Psi]$—	$[\Phi\diagdown\Psi]$—
126	8 сн.	$[\underbrace{[X\diagdown Y]/X}_{\gamma}]\,\underbrace{[Y/Y]}_{\delta}\,[\underbrace{[Y}_{\delta}\diagdown Y]/X]\,[Z/X]X$	$[\underbrace{[X\diagdown Y]/X]}_{\gamma}\,\underbrace{[Y/Y]}_{\delta}\,[\underbrace{[Y/Y]}_{\delta}\diagdown X]\,[Z/X]X$
127	4 сн.	$[\text{ПРЕДЛ}\diagdown V^{itr}_{\text{ж, ед, 3, }v}]$	$[\text{ПРЕДЛ}/V^{itr}_{\text{ж, ед, 3, }v}]$

А. Гладкий, И. Мельчук

Цена 61 коп.